Behavioral and Brain Sciences

cambridge.org/bbs

Meta-learned models of cognition

Marcel Binz[a,b], Ishita Dasgupta[c], Akshay K. Jagadish[a,b], Matthew Botvinick[c], Jane X. Wang[c] and Eric Schulz[a,b]

[a]Max Planck Institute for Biological Cybernetics, Tübingen, Germany; [b]Helmholtz Institute for Human-Centered AI, Munich, Germany and [c]Google DeepMind, London, UK
marcel.binz@helmholtz-munich.de
dasgupta.ishita@gmail.com
akshay.jagadish@tue.mpg.de
botvinick@google.com
wangjane@google.com
eric.schulz@tue.mpg.de

Target Article

Cite this article: Binz M, Dasgupta I, Jagadish AK, Botvinick M, Wang JX, Schulz E. (2024) Meta-learned models of cognition. *Behavioral and Brain Sciences* **47**, e147: 1–58. doi:10.1017/S0140525X23003266

Target Article Accepted: 19 November 2023
Target Article Manuscript Online: 23 November 2023
Commentaries Accepted: 13 February 2024

Keywords:
Bayesian inference; cognitive modeling; meta-learning; neural networks; rational analysis

What is Open Peer Commentary? What follows on these pages is known as a Treatment, in which a significant and controversial Target Article is published along with Commentaries (p. 20) and an Authors' Response (p. 53). See bbsonline.org for more information.

Corresponding author:
Marcel Binz;
Email: marcel.binz@helmholtz-munich.de

Abstract

Psychologists and neuroscientists extensively rely on computational models for studying and analyzing the human mind. Traditionally, such computational models have been hand-designed by expert researchers. Two prominent examples are cognitive architectures and Bayesian models of cognition. Although the former requires the specification of a fixed set of computational structures and a definition of how these structures interact with each other, the latter necessitates the commitment to a particular prior and a likelihood function that – in combination with Bayes' rule – determine the model's behavior. In recent years, a new framework has established itself as a promising tool for building models of human cognition: the framework of meta-learning. In contrast to the previously mentioned model classes, meta-learned models acquire their inductive biases from experience, that is, by repeatedly interacting with an environment. However, a coherent research program around meta-learned models of cognition is still missing to date. The purpose of this article is to synthesize previous work in this field and establish such a research program. We accomplish this by pointing out that meta-learning can be used to construct Bayes-optimal learning algorithms, allowing us to draw strong connections to the rational analysis of cognition. We then discuss several advantages of the meta-learning framework over traditional methods and reexamine prior work in the context of these new insights.

It is hard to imagine cognitive psychology and neuroscience without computational models – they are invaluable tools to study, analyze, and understand the human mind. Traditionally, such computational models have been hand-designed by expert researchers. In a cognitive architecture, for instance, researchers provide a fixed set of structures and a definition of how these structures interact with each other (Anderson, 2013b). In a Bayesian model of cognition, researchers instead specify a prior and a likelihood function that – in combination with Bayes' rule – fully determine the model's behavior (Griffiths, Kemp, & Tenenbaum, 2008). To provide one concrete example, consider the Bayesian model of function learning proposed by Lucas, Griffiths, Williams, and Kalish (2015). The goal of this model is to capture human learning in a setting that requires mapping input features to a numerical target value. When constructing their model, the authors had to hand-design a prior over functions that people expect to encounter. In this particular case, it was assumed that people prioritize linear functions over quadratic and other nonlinear functions.

The framework of meta-learning (Bengio, Bengio, & Cloutier, 1991; Schmidhuber, 1987; Thrun & Pratt, 1998) offers a radically different approach for constructing computational models by learning them through repeated interactions with an environment instead of requiring a priori specifications from a researcher. This process enables such models to acquire their inductive biases from experience, thereby departing from the traditional paradigm of hand-crafted models. For the function learning example mentioned above, this means that we do not need to specify which functions people expect to encounter in advance. Instead, during meta-learning a model would be exposed to many realistic function learning problems on which it then can figure out which functions are likely and which are not.

Recently, psychologists have started to apply meta-learning to the study of human learning (Griffiths et al., 2019). It has been shown that meta-learned models can capture a wide range of empirically observed phenomena that could not be explained otherwise. They, among others, reproduce human biases in probabilistic reasoning (Dasgupta, Schulz, Tenenbaum, & Gershman, 2020), discover heuristic decision-making strategies used by people (Binz, Gershman, Schulz, & Endres, 2022), and generalize compositionally on complex language tasks in a human-like manner (Lake & Baroni, 2023). The goal of the present article is to

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

develop a research program around meta-learned models of cognition and, in doing so, offer a synthesis of previous work and outline new research directions.

To establish such a research program, we will make use of a recent result from the machine learning community showing that meta-learning can be used to construct Bayes-optimal learning algorithms (Mikulik et al., 2020; Ortega et al., 2019; Rabinowitz, 2019). This correspondence is interesting from a psychological perspective because it allows us to connect meta-learning to another already well-established framework: the rational analysis of cognition (Anderson, 2013a; Chater & Oaksford, 1999). In a rational analysis, one first has to specify the goal of an agent along with a description of the environment the agent interacts with. The Bayes-optimal solution for the task at hand is then derived based on these assumptions and tested against empirical data. If needed, assumptions are modified and the whole process is repeated. This approach for constructing cognitive models has had a tremendous impact on psychology because it explains "why cognition works, by viewing it as an approximation to ideal statistical inference given the structure of natural tasks and environments" (Tenenbaum, 2021). The observation that meta-learned models can implement Bayesian inference implies that a meta-learned model can be used as a replacement for the corresponding Bayesian model in a rational analysis and thus suggests that any behavioral phenomenon that can be captured by a Bayesian model can also be captured by a meta-learned model.

Marcel Binz is a research scientist and deputy head of the Institute of Human-Centered AI at the Helmholtz Computational Health Center in Munich, and a guest scientist at the Max Planck Institute for Biological Cybernetics in Tübingen. He focuses on understanding human cognition from a computational perspective by utilizing tools from deep learning, reinforcement learning, Bayesian inference, and information theory. His work has been featured in top-tier journals such as *PNAS* and *Psychological Review*, as well as in leading machine learning venues such as NeurIPS and ICML.

Ishita Dasgupta is a senior research scientist at Google DeepMind. Her research is at the intersection of computational cognitive science and machine learning. She uses advances in machine learning to build new models of human reasoning, applies cognitive science approaches toward understanding black-box AI systems, and combines these insights to build more human-like artificial intelligence.

Akshay K. Jagadish is a PhD student jointly affiliated with the Max Planck Institute for Biological Cybernetics, Tübingen, and the Institute of Human-Centered AI at the Helmholtz Computational Health Center in Munich. His current research is dedicated to understanding the components essential for explaining human adaptive behavior across multiple task domains.

Matthew Botvinick, Senior Director of Research at Google DeepMind and Honorary Professor at Gatsby Unit for Computational Neuroscience, University College London, is the author of over 140 peer-reviewed publications in the areas of artificial intelligence, neuroscience, and cognitive psychology.

Jane X. Wang is a staff research scientist at Google DeepMind, where she researches meta-reinforcement learning, causal reasoning, and cognitively inspired AI. Her work has been published in various journals such as *Science*, *Neuron*, *Nature Neuroscience*, and top machine learning conferences such as NeurIPS, ICLR, and ICML.

Eric Schulz is the director of the Institute of Human-Centered AI at the Helmholtz Computational Health Center in Munich. He received his Ph.D. in Experimental Psychology from University College London in 2017. He is a recipient of the Robert J. Glushko Prize for Outstanding Doctoral Dissertation in Cognitive Science and a Jacobs Research Fellowship. His research focuses on understanding and improving human and machine learning.

We start our article by presenting a simplified version of an argument originally formulated by Ortega et al. (2019) and thereby make their result accessible to a broader audience. Having established that meta-learning produces models that can simulate Bayesian inference, we go on to discuss what additional explanatory power the meta-learning framework offers. After all, why should one not just stick to the tried-and-tested Bayesian approach? We answer this question by providing four original arguments in favor of the meta-learning framework (see Fig. 1 for a visual synopsis):

- Meta-learning can produce approximately optimal learning algorithms even if exact Bayesian inference is computationally intractable.
- Meta-learning can produce approximately optimal learning algorithms even if it is not possible to phrase the corresponding inference problem in the first place.
- Meta-learning makes it easy to manipulate a learning algorithm's complexity and can therefore be used to construct resource-rational models of learning.
- Meta-learning allows us to integrate neuroscientific insights into the rational analysis of cognition by incorporating these insights into model architectures.

The first two points highlight situations in which meta-learned models can be used for rational analysis but traditional Bayesian models cannot. The latter two points provide examples of how meta-learning enables us to make rational models of cognition more realistic, either by incorporating limited computational resources or neuroscientific insights. Taken together, these arguments showcase that meta-learning considerably extends the scope of rational analysis and thereby of cognitive theories more generally.

We will discuss each of these four points in detail and provide illustrations to highlight their relevance. We then reexamine prior studies from psychology and neuroscience that have applied meta-learning and put them into the context of our newly acquired insights. For each of the reviewed studies, we highlight how it relates to the four presented arguments, and discuss why its findings could not have been obtained using a classical Bayesian model. Following that, we describe under which conditions traditional models are preferable to those obtained by meta-learning. We finish our article by speculating what the future holds for meta-learning. Therein, we focus on how meta-learning could be the key to building a domain-general model of human cognition.

1. Meta-learned rationality

The prefix *meta-* is generally used in a self-referential sense: A meta-rule is a rule about rules, a meta-discussion is a discussion about discussions, and so forth. Meta-learning, consequently, refers to learning about learning. We, therefore, need to first establish a common definition of *learning* before covering meta-learning in more detail. For the present article, we adopt the following definition from Mitchell (1997):

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

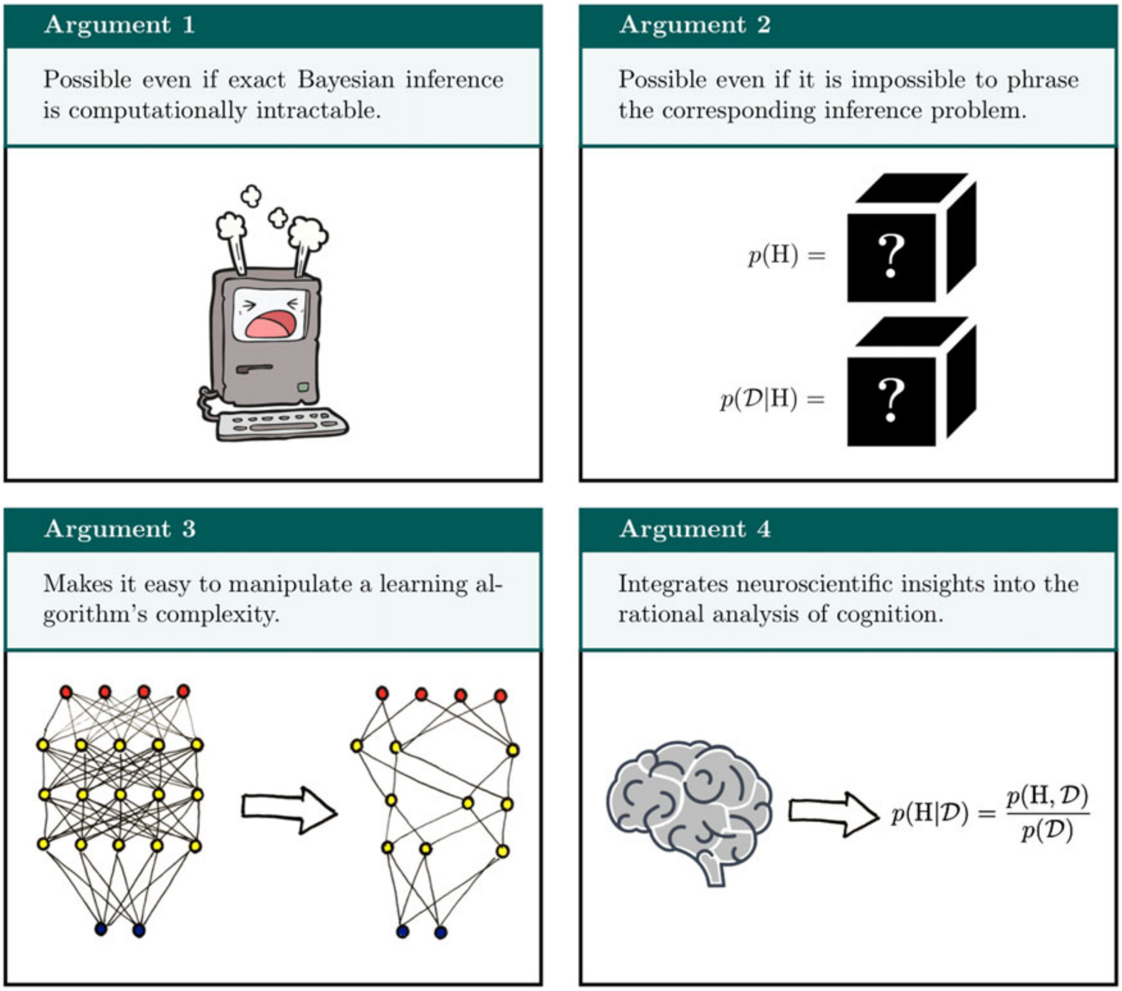

Argument 1
Possible even if exact Bayesian inference is computationally intractable.
Argument 2
Possible even if it is impossible to phrase the corresponding inference problem.
$p(\mathrm{H}) =$
?
$p(\mathcal{D}|\mathrm{H}) =$
?
Argument 3
Makes it easy to manipulate a learning algorithm's complexity.
Argument 4
Integrates neuroscientific insights into the rational analysis of cognition.
$p(\mathrm{H}|\mathcal{D}) = \frac{p(\mathrm{H}, \mathcal{D})}{p(\mathcal{D})}$

Figure 1. Visual synopsis of the four different arguments for meta-learning over Bayesian inference put forward in this article.

Definition: Learning
For a given task, training experience, and performance measure, an algorithm is said to learn if its performance at the task improves with experience.

To illustrate this definition, consider the following example which we will return to throughout the text: You are a biologist who has just discovered a new insect species and now set yourself the task of predicting how large members of this species are. You have already observed three exemplars in the wild with lengths of 16, 12, and 15 cm, respectively. These data amount to your training experience. Ideally, you can use this experience to make better predictions about the length of the next individual you encounter. You are said to have learned something if your performance is better after seeing the data than it was before. Typical performance measures for this example problem include the mean-squared error or the (negative) log-likelihood.

1.1 Bayesian inference for rational analyses

In a rational analysis of cognition, researchers are trying to compare human behavior to that of an optimal learning algorithm. However, it turns out that no learning algorithm is better than another when averaged over all possible problems (Wolpert, 1996; Wolpert & Macready, 1997), which means that we first have to make additional assumptions about the to-be-solved problem to obtain a well-defined notion of optimality. For our running example, one may make the following – somewhat unrealistic – assumptions:

(1) Each observed insect length x_k is sampled from a normal distribution with mean μ and standard deviation σ.
(2) An insect species' mean length μ cannot be observed directly, but the standard deviation σ is known to be 2 cm.
(3) Mean lengths across all insect species are distributed according to a normal distribution with a mean of 10 cm and a standard deviation of 3 cm.

An optimal way of making predictions about new observations under such assumptions is specified by Bayesian inference. Bayesian inference requires access to a prior distribution $p(\mu)$ that defines an agent's initial beliefs about possible parameter values before observing any data and a likelihood $p(x_{1:t}|\mu)$ that captures the agent's knowledge about how data are generated for a given set of parameters. In our running example, the prior and the likelihood can be identified as follows:

$$p(\mu) = N(\mu; 10, 3) \tag{1}$$

$$p(x_{1:t}|\mu) = \prod_{k=1}^{t} p(x_k|\mu) = \prod_{k=1}^{t} N(x_k; \mu, 2) \tag{2}$$

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

where $x_{1:t} = x_1, x_2, \ldots, x_t$ denote a sequence of observed insect lengths and the product in Equation (2) arises because of the additional assumption that observations are independent given the parameters.

The outcome of Bayesian inference is a posterior predictive distribution $p(x_{t+1}|x_{1:\,t})$, which the agent can use to make probabilistic predictions about a hypothetical future observation. To obtain this posterior predictive distribution, the agent first combines prior and likelihood into a posterior distribution over parameters by applying Bayes' theorem:

$$p\,(\mu|x_{1:t}) = \frac{p(x_{1:t}|\mu)p(\mu)}{\int p(x_{1:t}|\mu)p(\mu)d\mu} \tag{3}$$

In a subsequent step, the agent then averages over all possible parameter values weighted by their posterior probability to get the posterior predictive distribution:

$$p\,(x_{t+1}|x_{1:t}) = \int p(x_{t+1}|\mu)p(\mu|x_{1:t})d\mu \tag{4}$$

Multiple arguments justify Bayesian inference as a normative procedure, and thereby its use for rational analyses (Corner & Hahn, 2013). This includes Dutch book arguments (Lewis, 1999; Rescorla, 2020), free-energy minimization (Friston, 2010; Hinton & Van Camp, 1993), and performance-based justifications (Aitchison, 1975; Rosenkrantz, 1992). For this article, we are mainly interested in the latter class of performance-based justifications because these can be used – as we will demonstrate later on – to derive meta-learning algorithms that learn approximations to Bayesian inference.

Performance-based justifications are based on the notion of frequentist statistics. They assert that no learning algorithm can be better than Bayesian inference on a certain performance measure. Particularly relevant for this article is a theorem first proved by Aitchison (1975). It states that the posterior predictive distribution is the distribution from the set of all possible distributions Q that maximizes the log-likelihood of hypothetical future observations when averaged over the data-generating distribution $p\,(\mu,\, x_{1:t+1}) = p(\mu)p(x_{1:t+1}|\mu)$:

$$p\,(x_{t+1}|x_{1:t}) = \text{argmax}_{q \in Q} E_{p(\mu, x_{1:t+1})}[\log q\,(x_{t+1}|x_{1:t})] \tag{5}$$

Equation (5) implies that if an agent wants to make a prediction about the length of a still unobserved exemplar from a particular insect species and measures its performance using the log-likelihood, then – averaged across all possible species that can be encountered – there is no better way of doing it than using the posterior predictive distribution. We decided to include a short proof of this theorem within Box 1 for the curious reader as it does not appear in popular textbooks on probabilistic machine learning (Bishop, 2006; Murphy, 2012) nor in survey articles on Bayesian models of cognition. Note that, although the theorem itself is central to our later argument, working through its proof is not required to follow the remainder of this article.

Box 1 Proof: meta-learning maximizes log-likelihoods of future observations

We proof that the posterior predictive distribution $p(x_{t+1}|x_{1:t})$ maximizes the log-likelihood of future observations averaged over the data-generating distribution:

$$p(x_{t+1}|x_{1:t}) = \text{argmax}_q E_{p(\mu, x_{1:t+1})}[\log q(x_{t+1}|x_{1:t})] \tag{8}$$

The essence of this proof is to show that the posterior predictive distribution is superior to any other reference distribution $r(x_{t+1}|x_{1:t})$ in terms of log-likelihood:

$$E_{p(\mu, x_{1:t})}[\log p(x_{t+1}|x_{1:t})] \geq E_{p(\mu, x_{1:t})}[\log r(x_{t+1}|x_{1:t})]$$

or equivalently that:

$$E_{p(\mu, x_{1:t})}\left[\log \frac{p(x_{t+1}|x_{1:t})}{r(x_{t+1}|x_{1:t})}\right] \geq 0$$

Proofing this conjecture is straight-forward (Aitchison, 1975):

$$\begin{aligned}
&E_{p(\mu, x_{1:t})}\left[\log \frac{p(x_{t+1}|x_{1:t})}{r(x_{t+1}|x_{1:t})}\right] \\
&\quad = \sum\nolimits_{\mu} \sum\nolimits_{x_{1:t}} \sum\nolimits_{x_{t+1}} \log \frac{p(x_{t+1}|x_{1:t})}{r(x_{t+1}|x_{1:t})} p(x_{t+1}|\mu)p(x_{1:t}|\mu)p(\mu) \\
&\quad = \sum\nolimits_{x_{1:t}} \sum\nolimits_{\mu} \sum\nolimits_{x_{t+1}} \log \frac{p(x_{t+1}|x_{1:t})}{r(x_{t+1}|x_{1:t})} p(x_{t+1}|\mu)p(x_{1:t}|\mu)p(\mu) \\
&\quad = \sum\nolimits_{x_{1:t}} \sum\nolimits_{\mu} \sum\nolimits_{x_{t+1}} \log \frac{p(x_{t+1}|x_{1:t})}{r(x_{t+1}|x_{1:t})} p(x_{t+1}|\mu)p(\mu|x_{1:t})p(x_{1:t}) \\
&\quad = \sum\nolimits_{x_{1:t}} \left[\sum\nolimits_{\mu} \sum\nolimits_{x_{t+1}} \log \frac{p(x_{t+1}|x_{1:t})}{r(x_{t+1}|x_{1:t})} p(x_{t+1}|\mu)p(\mu|x_{1:t})\right] p(x_{1:t}) \\
&\quad = \sum\nolimits_{x_{1:t}} \left[\sum\nolimits_{x_{t+1}} \sum\nolimits_{\mu} \log \frac{p(x_{t+1}|x_{1:t})}{r(x_{t+1}|x_{1:t})} p(x_{t+1}|\mu)p(\mu|x_{1:t})\right] p(x_{1:t}) \\
&\quad = \sum\nolimits_{x_{1:t}} \left[\sum\nolimits_{x_{t+1}} \log \frac{p(x_{t+1}|x_{1:t})}{r(x_{t+1}|x_{1:t})} \left[\sum p(x_{t+1}|\mu)p(\mu|x_{1:t})\right]\right] p(x_{1:t}) \\
&\quad = \sum\nolimits_{x_{1:t}} \left[\sum\nolimits_{x_{t+1}} \log \frac{p(x_{t+1}|x_{1:t})}{r(x_{t+1}|x_{1:t})} p(x_{t+1}|x_{1:t})\right] p(x_{1:t}) \\
&\quad = \sum\nolimits_{x_{1:t}} KL\left[p(x_{t+1}|x_{1:t})|r(x_{t+1}|x_{1:t})\right] p(x_{1:t}) \\
&\quad \geq 0
\end{aligned}$$

Note that although we used sums in our proof, thereby assuming that relevant quantities take discrete values, the same ideas can be readily applied to continuous-valued quantities by replacing sums with integrals.

1.2 Meta-learning

Having summarized the general concepts behind Bayes-optimal learning, we can now start to describe meta-learning in more detail. Formally speaking, a meta-learning algorithm is defined as any algorithm that "uses its experience to change certain aspects of a learning algorithm, or the learning method itself, such that the modified learner is better than the original learner at learning from additional experience" (Schaul & Schmidhuber, 2010).

To accomplish this, one first decides on an inner-loop (or base) learning algorithm and determines which of its aspects can be modified. We also refer to these modifiable aspects as meta-parameters (i.e., meta-parameters are simply parameters of a system that are adapted during meta-learning). In an outer-loop (or meta-learning) process, the system is then trained on a series of learning problems such that the inner-loop learning algorithm gets better at solving the problems that it encounters. We provide a high-level overview of this framework in Figure 2.

The previous definition is quite broad and includes a variety of methods. It is, for example, possible to meta-learn:

- Hyperparameters for a base learning algorithm, such as learning rates, batch sizes, or the number of training epochs

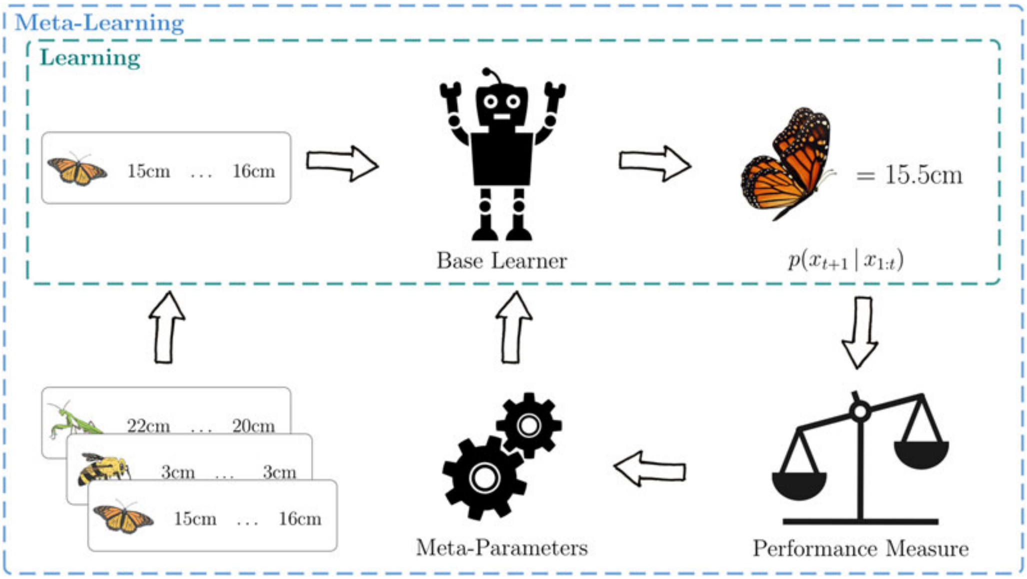


Figure 2. High-level overview of the meta-learning process. A base learner (green rectangle) receives data and performs some internal computations that improve its predictions on future data-points. A meta-learner (blue rectangle) encompasses a set of meta-parameters that can be adapted to create an improved learner. This is accomplished by training the learner on a distribution of related learning problems.

(Doya, 2002; Feurer & Hutter, 2019; Li, Zhou, Chen, & Li, 2017).

- Initial parameters of a neural network that is trained via stochastic gradient descent (Finn, Abbeel, & Levine, 2017; Nichol, Achiam, & Schulman, 2018).
- Prior distributions in a probabilistic graphical model (Baxter, 1998; Grant, Finn, Levine, Darrell, & Griffiths, 2018).
- Entire learning algorithms (Hochreiter, Younger, & Conwell, 2001; Santoro, Bartunov, Botvinick, Wierstra, & Lillicrap, 2016).

Although all these methods have their own merits, we will be primarily concerned with the latter approach. Learning entire learning algorithms from scratch is arguably the most general and ambitious type of meta-learning, and it is the focus of this article because it is the only one among the aforementioned approaches leading to Bayes-optimal learning algorithms that can be used for rational analyses.

1.3 Meta-learned inference

It may seem like a daunting goal to learn an entire learning algorithm from scratch, but the core idea behind the approach we discuss in the following is surprisingly simple: Instead of using Bayesian inference to obtain the posterior predictive distribution, we teach a general-purpose function approximator to do this inference. Previous work has mostly focused on using recurrent neural networks as function approximators in this setting and thus we will – without loss of generality – focus our upcoming exposition on this class of models.

Like the posterior predictive distribution, the recurrent neural network processes a sequence of observed length from a particular insect species and produces a predictive distribution over the lengths of potential future observations from the same species. More concretely, the meta-learned predictive distribution takes a predetermined functional form whose parameters are given by the network outputs. If we had, for example, decided to use a normal distribution as the functional form of the meta-learned predictive distribution, outputs of the network would correspond to an expected length m_{t+1} and its standard deviation s_{t+1}. Figure 3a illustrates this setup graphically.

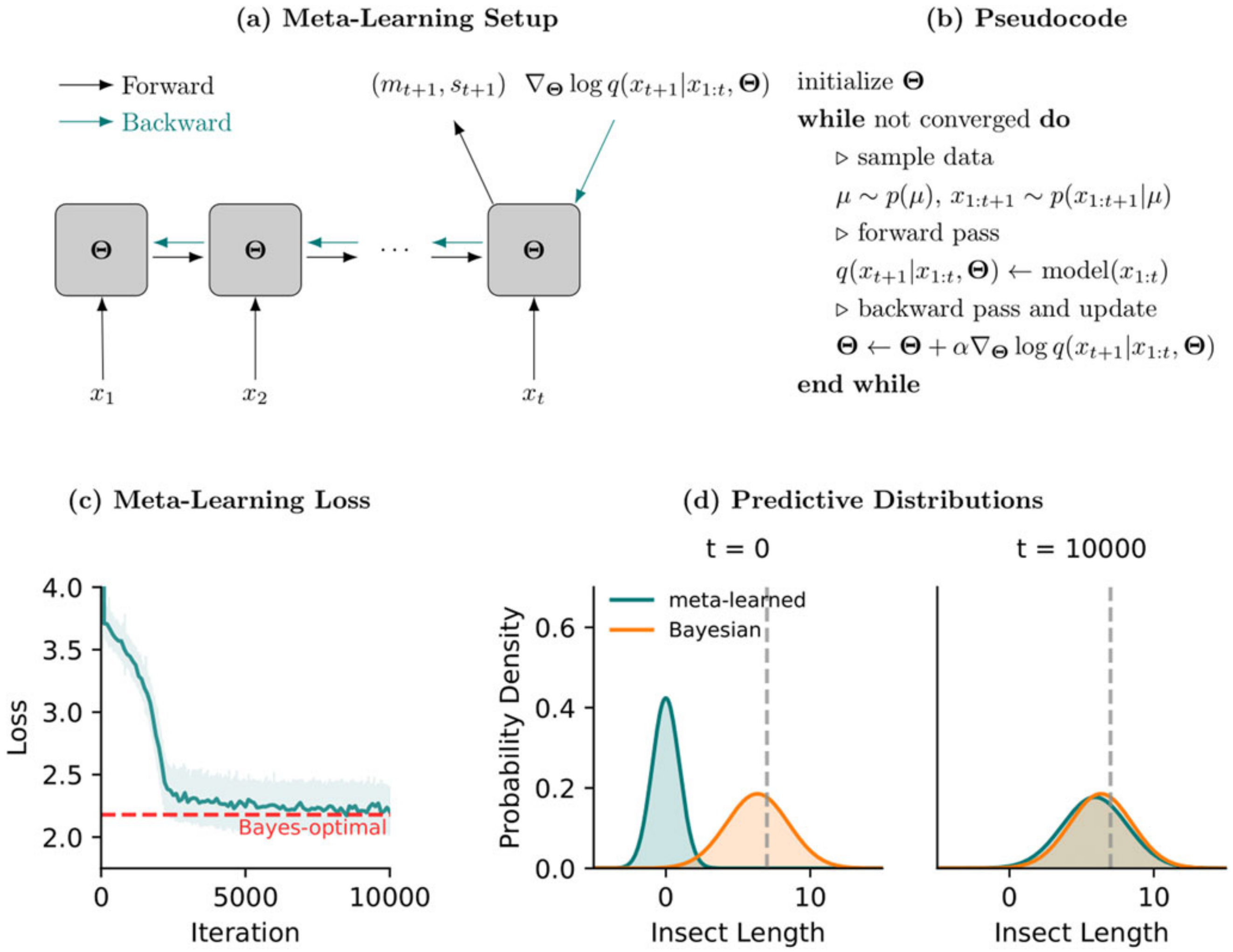
Initially, the recurrent neural network implements a randomly initialized learning algorithm.[1] The goal of the meta-learning process is then to turn this system into an improved learning algorithm. The final result is a learning algorithm that is *learned* or trained rather than specified by a practitioner. To create a learning signal to do this training, we need a performance measure that can be used to optimize the network. Equation (5) suggests a straightforward strategy for designing such a measure by replacing the maximization over all possible distributions with a maximization over meta-parameters Θ (in our case, the weights of the recurrent neural network):

$$\begin{aligned} &\text{argmax}_{q \in Q} E_{p(\mu, x_{1:t+1})}[\log q(x_{t+1} | x_{1:t})] \\ &\quad \approx \text{argmax}_{\Theta} E_{p(\mu, x_{1:t+1})}[\log q(x_{t+1} | x_{1:t}, \Theta)] \end{aligned} \tag{6}$$

To turn this expression into a practical meta-learning algorithm, we will – as common practice when training deep neural networks – maximize a sample-based version using stochastic gradient ascent:

$$\begin{aligned} &\text{argmax}_{\Theta} E_{p(\mu, x_{1:t+1})}[\log q(x_{t+1} | x_{1:t}, \Theta)] \\ &\quad \approx \text{argmax}_{\Theta} \frac{1}{N} \sum_{n=1}^{N} \log q(x_{t+1}^{(n)} | x_{1:t}^{(n)}, \Theta) \end{aligned} \tag{7}$$

Figure 3b presents pseudocode for a simple gradient-based procedure that maximizes Equation (7). The entire meta-learning algorithm can be implemented in just around 30 lines of self-contained *PyTorch* code (Paszke et al., 2019). We provide an annotated reference implementation on this article's accompanying Github repository.[2]



```
initialize $\Theta$
while not converged do
    ▷ sample data
    $\mu \sim p(\mu)$, $x_{1:t+1} \sim p(x_{1:t+1}|\mu)$
    ▷ forward pass
    $q(x_{t+1}|x_{1:t}, \Theta) \leftarrow \text{model}(x_{1:t})$
    ▷ backward pass and update
    $\Theta \leftarrow \Theta + \alpha \nabla_{\Theta} \log q(x_{t+1}|x_{1:t}, \Theta)$
end while
```

Figure 3. Meta-learning illustration. (a) A recurrent neural network processes a sequence of observations and produces a predictive distribution at the final time-step. (b) Pseudocode for a simple meta-learning algorithm. (c) Loss during meta-learning with shaded contours corresponding to the standard deviation across 30 runs. (d) Posterior and meta-learned predictive distributions for an example sequence at beginning and end of meta-learning. The dotted gray line denotes the (unobserved) mean length.

1.4 How good is a meta-learned algorithm?

We have previously shown that the global optimum of Equation (7) is achieved by the posterior predictive distribution. Thus, by maximizing this performance measure, the network is actively encouraged to implement an approximation to exact Bayesian inference. Importantly, after meta-learning is completed, producing an approximation to the posterior predictive distribution does not require any further updates to the network weights. To perform an inference (i.e., the learn), we simply have to query the network's outputs after providing it with a particular sequence of observations. Learning at this stage is then realized by updating the hidden activations of the recurrent neural network as opposed to its weights. The characteristics of this new activation-based learning algorithm can be potentially vastly different from the weight-based learning algorithm used for meta-learning.

If we want to use the fully optimized network for rational analyses, we have to ask ourselves: How well does the resulting model approximate Bayesian inference? Two aspects have to be considered when answering this question. First, the network has to be sufficiently expressive to produce the exact posterior predictive distribution for all input sequences. Neural networks of sufficient width are universal function approximators (Hornik, Stinchcombe, & White, 1989), meaning that they can approximate any continuous function to arbitrary precision. Therefore, this aspect is not too problematic for the optimality argument. The second aspect is a bit more intricate: Assuming that the network is powerful enough to represent the global optimum of Equation (7), the employed optimization procedure also has to find it. Although we are not aware of any theorem that could provide such a guarantee, in practice, it has been observed that meta-learning procedures similar to the one discussed here often lead to networks that closely approximate Bayesian inference (Mikulik et al., 2020; Rabinowitz, 2019). We provide a visualization demonstrating that the predictions of a meta-learned model closely resemble those of exact Bayesian inference for our insect length example in Figures 3c and 3d.

Although our exposition in this section focused on the supervised learning case, the same ideas can also be readily extended to the reinforcement learning setting (Duan et al., 2016; Wang et al., 2016). Box 2 outlines the general ideas behind the meta-reinforcement learning framework.

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

Box 2 Meta-reinforcement learning

The main text has focused on tasks in which an agent receives direct feedback about which response would have been correct. In the real world, however, people do not always receive such explicit feedback. They, instead, often have to deal with partial information – taking the form of rewards, utilities, or costs – that merely informs them about the quality of their response.

Problems that fall into this category are often modeled as Markov decision processes (MDPs). In an MDP, an agent repeatedly interacts with an environment. In each time-step, it observes the state of the environment s_t and can take an action a_t that leads to a reward signal r_t sampled from a reward distribution $p(r_t|s_t, a_t, \mu_r)$. Executing an action furthermore influences the environment state at the next time-step according to a transition distribution $p(s_{t+1}|s_t, a_t, \mu_s)$.

The goal of a Bayes-optimal reinforcement learning agent is to find a policy, which is a mapping from a history of observations $h_t = s_1, a_1, r_1, \ldots, s_{t-1}, a_{t-1}, r_{t-1}, s_t$ to a probability distribution over actions $\pi^*(a_t|h_t)$, that maximizes the total amount of obtained rewards across a finite horizon H averaged over all problems that may be encountered:

$$\pi^*(a_t|h_t) = \text{argmax}_\pi E_{p(\mu_r,\mu_s)\prod p(r_t|s_t,a_t,\mu_r)p(s_{t+1}|s_t,a_t,\mu_s)\pi(a_t|h_t)}\left[\sum_{t=1}^{H} r_t\right] \quad (9)$$

MDPs with unknown reward and transition distributions are substantially more challenging to solve optimally compared to supervised problems as there is no teacher informing the agent about which actions are right or wrong. Instead, the agent has to figure out the most rewarding course of action solely through trial and error. Finding an analytical solution to Equation (9) is extremely challenging and indeed only possible for a few special cases (Duff, 2003; Gittins, 1979), which made it historically near impossible to investigate such problems within the framework of rational analysis.

Even though finding an analytical expression of the Bayes-optimal policy is often impossible, it is straightforward to meta-learn an approximation to it (Duan et al., 2016; Wang et al., 2016). The general concept is almost identical to the supervised learning case: Parameterize the to-be-learned policy with a recurrent neural network and replace the maximization over the set of all possible policies from Equation (9) with a sample-based approximation that maximizes over neural network parameters. The resulting problem can then be solved using any standard deep reinforcement learning algorithm.

Like in the supervised learning case, the resulting recurrent neural network implements a free-standing reinforcement learning algorithm after meta-learning is completed. Learning is once again implemented via a simple forward pass through the network, i.e., by conditioning the model on an additional data-point. The meta-learned reinforcement learning algorithm approximates the Bayes-optimal policy under the same conditions as in the supervised learning case: A sufficiently expressive model and an optimization procedure that is able to find the global optimum.

1.5 Tool or theory?

It is often not so trivial to separate meta-learning from normal learning. We believe that part of this confusion arises from an underspecification regarding what is being studied. In particular, the meta-learning framework provides opportunities to address two distinct research questions:

(1) It can be used to study how people improve their learning abilities over time.
(2) It can be used as a methodological tool to construct learning algorithms with the properties of interest (and thereafter compare the emerging learning algorithms to human behavior).

Historically, behavioral psychologists have been mainly interested in the former aspect (Doya, 2002; Harlow, 1949). In the 1940s, for example, Harlow (1949) already studied how learning in monkeys improves over time. He found that they adapted their learning strategies after sufficiently many interactions with tasks that shared a common structure, thereby showing a learning-to-learn effect. By now, examples of this phenomenon have been found in many different species – including humans – across nature (Wang, 2021).

More recently, psychologists have started to view meta-learning as a methodological tool to construct approximations to Bayes-optimal learning algorithms (Binz et al., 2022; Kumar, Dasgupta, Cohen, Daw, & Griffiths, 2020a), and subsequently use the resulting algorithms to study human cognition. The key difference from the former approach is that, in this setting, one abstracts away from the process of meta-learning and instead focuses on its outcome. From this perspective, only the fully converged model is of interest. Importantly, this approach allows us to investigate human learning from a rational perspective because we have demonstrated that meta-learning can be used to construct approximations to Bayes-optimal learning.

We place an emphasis on the second aspect in the present article and advocate for using fully converged meta-learned algorithms – as replacements for the corresponding Bayesian models – for rational analyses of cognition.[3] In the next section, we will outline several arguments that support this approach. However, it is important to mention that we believe that meta-learning can also be a valuable tool to understand the process of learning-to-learn itself. In this context, several intriguing questions arise: At what timescale does meta-learning take place in humans? How much of it is because of task-specific adaptations? How much of it is based on evolutionary or developmental processes? Although we agree that these are important questions, they are not the focus of this article.

2. Why not Bayesian inference?

We have just argued that it is possible to meta-learn Bayes-optimal learning algorithms. What are the implications of this result? If one has access to two different theories that make identical predictions, which of them should be preferred? Bayesian inference has already established itself as a valuable tool for building cognitive models in the recent decades. Thus, the burden of proof is arguably on the meta-learning framework. In this section, we provide four different arguments that highlight the advantages of meta-learning for building models of cognition. Many of these arguments are novel and have not been put forward explicitly in previous literature. The first two arguments highlight situations in which meta-learned models can be used for rational analysis but traditional Bayesian models cannot. The latter two provide examples of how meta-learning enables us to make rational models of cognition more realistic, either by incorporating limited computational resources or neuroscientific insights.

2.1 Intractable inference

> **Argument 1**
> Meta-learning can produce approximately optimal learning algorithms even if exact Bayesian inference is computationally intractable.

Bayesian inference becomes intractable very quickly because the complexity of computing the normalization constant that appears in the denominator grows exponentially with the number of unobserved parameters. In addition, it is only possible to find a closed-form expression of the posterior distribution for certain combinations of prior and likelihood. In our running example, we assumed that both prior and likelihood follow a normal distribution, which, in turn, leads to a normally distributed posterior. However, if one would instead assume that the prior over mean

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

length follows an exponential distribution – which is arguably a more sensible assumption as it enforces lengths to be positive – it becomes already impossible to find an analytical expression for the posterior distribution.

Researchers across disciplines have recognized these challenges and have, in turn, developed approaches that can approximate Bayesian inference without running into computational difficulties. Prime examples of this are variational inference (Jordan, Ghahramani, Jaakkola, & Saul, 1999) and Markov chain Monte-Carlo (MCMC) methods (Geman & Geman, 1984). In variational inference, one phrases inference as an optimization problem by positing a variational approximation whose parameters are fitted to minimize a divergence measure to the true posterior distribution. MCMC methods, on the other hand, draw samples from a Markov chain that has the posterior distribution as its equilibrium distribution. Previous research in cognitive science indicates that human learning shows characteristics of such approximations (Courville & Daw, 2008; Dasgupta, Schulz, & Gershman, 2017; Daw, Courville, & Dayan, 2008; Sanborn, Griffiths, & Navarro, 2010; Sanborn & Silva, 2013).

Meta-learned inference also never requires an explicit calculation of the exact posterior or posterior predictive distribution. Instead, it performs approximately optimal inference via a single forward pass through the network. Inference, in this case, is approximate because we had to determine a functional form for the predictive distribution. The chosen form may deviate from the true form of the posterior predictive distribution, which, in turn, leads to approximation errors.[4] In some sense, this type of approximation is similar to variational inference: Both approaches involve optimization and require one to define a functional form of the respective distribution. However, the optimization process in both approaches uses a different loss function and happens at different timescales. Although variational inference employs the negative evidence lower bound as its loss function, meta-learning directly maximizes for models that can be expected to generalize well to unseen observations (using the performance-based measure from Equation (5)). Furthermore, meta-learned inference only involves optimization during the outer-loop meta-learning process but not during the actual learning itself. To update how a meta-learned model makes predictions in the light of new data, we only have to perform a simple forward pass through the network. In contrast to this, standard variational inference requires us to rerun the whole optimization process from scratch every time a new data-point is observed.[5]

In summary, it is possible to meta-learn an approximately Bayes-optimal learning algorithm. If exact Bayesian inference is not tractable, such models are our best option for performing rational analyses. Yet, many other methods for approximate inference, such as variational inference and MCMC methods, also share this feature, and it will thus ultimately be an empirical question which of these approximations provides a better description of human learning.

2.2 Unspecified problems

Argument 2

Meta-learning can produce optimal learning algorithms even if it is not possible to phrase the corresponding inference problem in the first place.

Bayesian inference is hard, but posing the correct inference problem can be even harder. What exactly do we mean by that? To perform Bayesian inference, we need to specify a prior and a likelihood. Together, these two objects fully specify the assumed data-generating distribution, and thus the inference problem. Ideally, the specified data-generating distribution should match how the environment actually generates its data. It is fairly straightforward to fulfill this requirement in artificial scenarios, but for many real-world problems, it is not. Take for instance our running example: Does the prior over mean length really follow a normal distribution? If yes, what are the mean and variance of this distribution? Are the underlying parameters actually time-invariant or do they, for example, change based on seasons? None of these questions can be answered with certainty.

In his seminal work on Bayesian decision theory, Savage (1972) made the distinction between small- and large-world problems. A small-world problem is one "in which all relevant alternatives, their consequences, and probabilities are known" (Gigerenzer & Gaissmaier, 2011). A large-world problem, on the other hand, is one in which the prior, the likelihood, or both cannot be identified. Savage's distinction between small and large worlds is relevant for the rational analysis of human cognition as its critics have pointed out that Bayesian inference only provides a justification for optimal reasoning in small-world problems (Binmore, 2007) and that "very few problems of interest to the cognitive, behavioral, and social sciences can be said to satisfy [this] condition" (Brighton & Gigerenzer, 2012).

Identifying the correct set of assumptions becomes especially challenging once we deal with more complex problems. To illustrate this, consider a study conducted by Lucas et al. (2015) who attempted to construct a Bayesian model of human function learning. Doing so required them to specify a prior over functions that people expect to encounter. Without direct access to such a distribution, they instead opted for a heuristic solution: 98.8% of functions are expected to be linear, 1.1% are expected to be quadratic, and 0.1% are expected to be nonlinear. Empirically, this choice led to good results, but it is hard to justify from a rational perspective. We simply do not know the frequency with which these functions appear in the real world, nor whether the given selection fully covers the set of functions expected by participants.

There are also inference problems in which it is not possible to specify or compute the likelihood function. These problems have been studied extensively in the machine learning community under the names of simulation-based or likelihood-free inference (Cranmer, Brehmer, & Louppe, 2020; Lueckmann, Boelts, Greenberg, Goncalves, & Macke, 2021). In this setting, it is typically assumed that we can sample data from the likelihood for a given parameter setting but that computing the corresponding likelihood is impossible. Take, for instance, our insect length example. It should be clear that an insect's length does not only depend on its species' mean but also on many other factors such as climate, genetics, and the individual's age. Even if all these factors were known, mapping them to a likelihood function does seem close to impossible.[6] But, we can generate samples easily by observing insects in the wild. If we had access to large database of insect length measurements for different species, this could be directly used to meta-learn an approximately Bayes-optimal learning algorithm for predicting their length, while circumventing an explicit definition of a likelihood function.

In cases where we do not have access to a prior or a likelihood, we can neither apply exact Bayesian inference nor approximate inference schemes such as variational inference or MCMC methods. In contrast to this, meta-learned inference does not require us to define the prior or the likelihood explicitly. It only demands samples from the data-generating distribution to meta-learn an approximately Bayes-optimal learning algorithm – a much weaker requirement (Müller, Hollmann, Arango, Grabocka, & Hutter, 2021). The ability to construct Bayes-optimal learning algorithms for large-world problems is a unique feature of the meta-learning framework, and we believe that it could open up totally new avenues for constructing rational models of human cognition. To highlight one concrete example, it would be possible to take a collection of real-world decision-making tasks – such as the ones presented by Czerlinski et al. (1999) – and use them to obtain a meta-learned agent that is adapted to the decision-making problems that people actually encounter in their everyday lives. This algorithm could then serve as a normative standard against which we can compare human decision making.

2.3 Resource rationality

> **Argument 3**
> Meta-learning makes it easy to manipulate a learning algorithm's complexity and can therefore be used to construct resource-rational models of learning.

Bayesian inference has been successfully applied to model human behavior across a number of domains, including perception (Knill & Richards, 1996), motor control (Körding & Wolpert, 2004), everyday judgments (Griffiths & Tenenbaum, 2006), and logical reasoning (Oaksford et al., 2007). Notwithstanding these success stories, there are also well-documented deviations from the notion of optimality prescribed by Bayesian inference. People, for example, underreact to prior information (Kahneman & Tversky, 1973), ignore evidence (Benjamin, 2019), and rely on heuristic decision-making strategies (Gigerenzer & Gaissmaier, 2011).

The intractability of Bayesian inference – together with empirically observed deviations from it – has led researchers to conjecture that people only attempt to approximate Bayesian inference. Many different notions of what constitutes a computational resource have been suggested, such as memory (Dasgupta & Gershman, 2021), thinking time (Ratcliff & McKoon, 2008), or physical effort (Hoppe & Rothkopf, 2016).

Cover (1999) relies on a dichotomy that will be useful for our following discussion. He refers to the algorithmic complexity of an algorithm as the number of bits needed to *implement* it. In contrast, he refers to the computational complexity of an algorithm as the space, time, or effort required to *execute* it. It is possible to cast many approximate inference schemes as resource-rational algorithms (Sanborn, 2017). The resulting models typically consider some form of computational complexity. In MCMC methods, computational complexity can be measured in terms of the number of drawn samples: Drawing fewer samples leads to faster inference at the cost of introducing a bias (Courville & Daw, 2008; Sanborn et al., 2010). In variational inference, on the other hand, it is possible to introduce an additional parameter that allows to trade-off performance against the computational complexity of transforming the prior into the posterior distribution (Binz & Schulz, 2022b; Ortega, Braun, Dyer, Kim, & Tishby, 2015). Likewise, other frameworks for building resource-rational models, such as rational meta-reasoning (Lieder & Griffiths, 2017), also only target computational complexity.

The prevalence of resource-rational models based on computational complexity is likely because of the fact that building similar models based on algorithmic complexity is much harder. Measuring algorithmic complexity historically relies on the notion of Kolmogorov complexity, which is the size of the shortest computer program that produces a particular data sequence (Chaitin, 1969; Kolmogorov, 1965; Solomonoff, 1964). Kolmogorov complexity is in general noncomputable, and, therefore, of limited practical interest.[7]

Meta-learning provides us with a straightforward way to manipulate both algorithmic and computational complexity in a common framework by adapting the size of the underlying neural network model. Limiting the complexity of network weights places a constraint on algorithmic complexity (as reducing the number of weights decreases the number of bits needed to store them, and hence also the number of bits needed to store the learning algorithm). Limiting the complexity of activations, on the other hand, places a constraint on computational complexity (reducing the number of hidden units, e.g., decreases the memory needed for executing the meta-learned model). This connection can be made more formal in an information-theoretic framework (Hinton & Van Camp, 1993; Hinton & Zemel, 1993). For applications of this idea in the context of human cognition, see, for instance, Binz et al. (2022) or Bates and Jacobs (2020).

Previously, both forms of complexity constraints have been realized in meta-learned models. Dasgupta et al. (2020) decreased the number of hidden units of a meta-learned inference algorithm, effectively reducing its computational complexity. In contrast, Binz et al. (2022) placed a constraint on the description length of neural network weights (i.e., the number of bits required to store them), which implements a form of algorithmic complexity. To the best of our knowledge, no other class of resource-rational models exists that allows us to take both algorithmic and computational complexity into account, making this ability a unique feature of the meta-learning framework.

2.4 Neuroscience

> **Argument 4**
> Meta-learning allows us to integrate neuroscientific insights into the rational analysis of cognition by incorporating these insights into model architectures.

In addition to providing a framework for understanding many aspects of behavior, meta-learning offers a powerful lens through which to view brain structure and function. For instance, Wang et al. (2018) presented observations supporting the hypothesis that prefrontal circuits may constitute a meta-reinforcement learning system. From a computational perspective, meta-learning strives to learn a faster inner-loop learning algorithm via an adjustment of neural network weights in a slower outer-loop learning process. Within the brain, an analogous process plausibly occurs when slow, dopamine-driven synaptic change gives rise to reinforcement learning processes that occur within the activity dynamics of the prefrontal network, allowing for adaptation on much faster timescales. This perspective recontextualized the role of dopamine function in reward-based learning and was able to account for a range of previously puzzling neuroscientific findings. To highlight one example, Bromberg-Martin, Matsumoto, Hong, and Hikosaka (2010) found that dopamine signaling reflected updates in not only *experienced* but also *inferred* values of targets. Notably, a meta-reinforcement learning agent trained on the same task also recovered this pattern. Having a mapping of meta-reinforcement learning components onto existing brain regions furthermore allows us to apply experimental manipulations that directly perturb neural activity, for example by using optogenetic techniques. Wang et al. (2018) used this idea to

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

modify their original meta-reinforcement learning architecture to mimic the blocking or enhancement of dopaminergic reward prediction error signals, in direct analogy with optogenetic stimulation delivered to rats performing a two-armed bandit task (Stopper, Maric, Montes, Wiedman, & Floresco, 2014).

Importantly, the direction of exchange can also work in the other direction, with neuroscientific findings constraining and inspiring new forms of meta-learning architectures. Bellec, Salaj, Subramoney, Legenstein, and Maass (2018), for example, showed that recurrent networks of spiking neurons are able to display convincing learning-to-learn behavior, including in the realm of reinforcement learning. Episodic meta-reinforcement learning (Ritter et al., 2018) architectures are also heavily inspired by neuroscientific accounts of complementary learning systems in the brain (McClelland, McNaughton, & O'Reilly, 1995). Both of these examples demonstrate that meta-learning can be used to build more biologically plausible learning algorithms, and thereby highlight that it can act as a bridge between Marr's computational and implementational levels (Marr, 2010).

Finally, the meta-learning perspective not only allows us to connect machine learning and neuroscience via architectural design choices but also via the kinds of tasks that are of interest. Dobs, Martinez, Kell, and Kanwisher (2022), for instance, suggested that functional specialization in neural circuits, which has been widely observed in biological brains, arises as a consequence of task demands. In particular, they found that convolutional neural networks "optimized for both tasks spontaneously segregate themselves into separate systems for faces and objects." Likewise, Yang, Joglekar, Song, Newsome, and Wang (2019) found that training a single recurrent neural network to perform a wide range of cognitive tasks yielded units that were clustered along different functional cognitive processes. Put another way, it seems plausible that functional specialization emerges by training neural networks on multiple tasks. Although this has not been tested so far, we speculate that this also holds in the meta-learning setting, as it involves training on multiple tasks by design. If this were true, we could look at the emerging areas inside a meta-learned model, and use the resulting insights to generate novel predictions about the processes happening in individual brain areas (Kanwisher, Khosla, & Dobs, 2023).

3. Previous research

Meta-learned models are already starting to transform the cognitive sciences today. They allow us to model things that are hard to capture with traditional models such as compositional generalization, language understanding, and model-based reasoning. In this section, we provide an overview of what has been achieved with the help of meta-learning in previous work. We arranged this review into various thematic subcategories. For each of them, we summarize which key findings have been obtained by meta-learning and discuss why these results would have been difficult to obtain using traditional models of learning by appealing to the insights from the previous section.

3.1 Heuristics and cognitive biases

Meta-learning has been previously used to discover algorithms with a limited computational budget that show human-like cognitive biases as we have already alluded to earlier. Dasgupta et al. (2020) trained a neural network on a distribution of probabilistic inference problems while controlling for the number of its hidden units. They found that their model – when restricted to just a single hidden unit – captured many biases in human reasoning, including a conservatism bias and base rate neglect. Likewise, Binz et al. (2022) trained a neural network on a distribution of decision-making problems while controlling for the number of bits needed to represent the network. Their model discovered two previously suggested heuristics in specific environments and made precise prognoses about when these heuristics should be applied. In particular, knowing the correct ranking of features led to one reason decision making, knowing the directions of features led to an equal weighting heuristic, and not knowing about either of them led to strategies that use weighted combinations of features (also see Figs. 4a and 4b).

In both of these studies, meta-learned models offered a novel perspective on results that were previously viewed as contradictory. This was in part possible because meta-learning enabled us to easily manipulate the complexity of the underlying learning algorithm. Although doing so is, at least in theory, also possible within the Bayesian framework, no Bayesian model that captures the full set of findings from Dasgupta et al. (2020) and Binz et al. (2022) has been discovered so far. We hypothesize that this could be because traditional rational process models struggle to capture that human strategy selection is context-dependent even before receiving any direct feedback signal (Mercier & Sperber, 2017). The meta-learned models of Dasgupta et al. (2020) and Binz et al. (2022), on the other hand, were able to readily show context-specific biases when trained on an appropriate task distribution.

3.2 Language understanding

Meta-learning may also help us to answer questions regarding how people process, understand, and produce language. Whether the inductive biases needed to acquire a language are learned from experience or are inherited is one of these questions (Yang & Piantadosi, 2022). McCoy, Grant, Smolensky, Griffiths, and Linzen (2020) investigated how to equip a model with a set of linguistic inductive biases that are relevant to human cognition. Their solution to this problem builds upon the idea of model-agnostic meta-learning (Finn et al., 2017). In particular, they meta-learned the initial weights of a neural network such that the network can adapt itself quickly to new languages using standard gradient-based learning. When being trained on a distribution over languages, these initial weights can be interpreted as universal factors that are shared across all languages. They showed that this approach identifies inductive biases (e.g., a bias for treating certain phonemes as vowels) that are useful for acquiring a language's syllable structure. Although their current work makes limited claims about human language acquisition, their approach be used in future studies to disentangle which inductive biases are learned from experience and which ones are inherited. They additionally argued that a Bayesian modeling approach would only be able to consider a restrictive set of inductive biases as it needs to commit to a particular representation and inference algorithm. In contrast, the meta-learning framework made it easy to implement the intended inductive biases by simply manipulating the distribution of encountered languages.

The ability to compose simple elements into complex entities is at the heart of human language. The property of languages to "make infinite use of finite means" (Chomsky, 2014) is what allows us to make strong generalizations from limited data. For

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

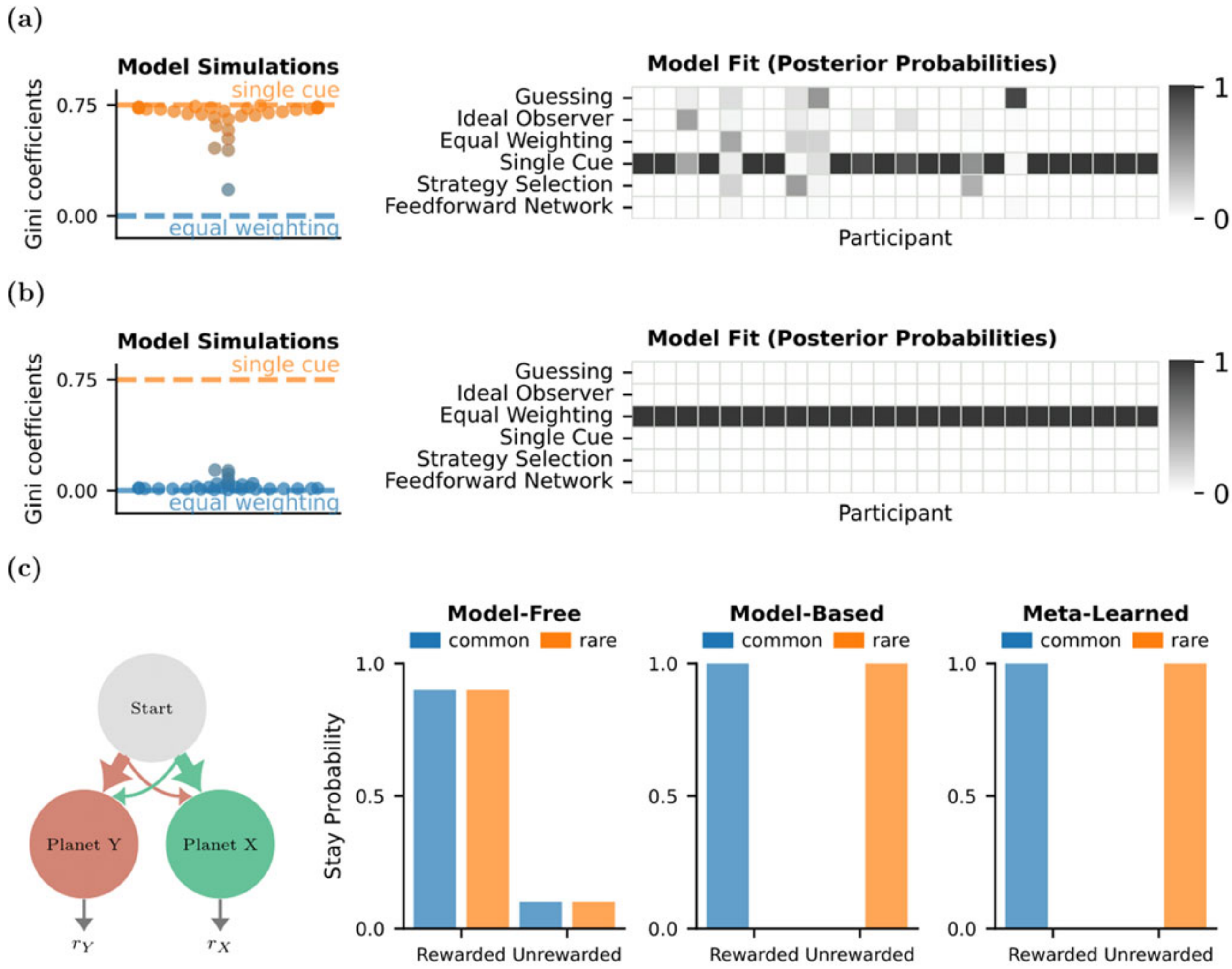


Figure 4. Example results obtained using meta-learned models. (a) In a paired comparison task, a meta-learned model identified a single-cue heuristic as the resource-rational solution when information about the feature ranking was available. Follow-up experiments revealed that people indeed apply this heuristic under the given circumstances. (b) If information about feature directions was available, the same meta-learned model identified an equal weighting heuristic as the resource-rational solution. People also applied this heuristic in the given context (Binz et al., 2022). (c) Wang et al. (2016) showed that meta-learned models can exhibit model-based learning characteristics in the two-step task (Daw et al., 2011) even when they were purely trained through model-free approaches. The plots on the right illustrate the probability of repeating the previous action for different agents (model-free, model-based, meta-learned) after a common or uncommon transition and after a received or omitted reward.

example, people readily understand what it means to "dax twice" or to "dax slowly" after learning about the meaning of the verb "dax." How to build models with a similar proficiency, however, remains an open research question. Lake (2019) showed that a transformer-like neural network can be trained to make such compositional generalizations through meta-learning. Importantly, during meta-learning, his models were adapted to problems that required compositional generalization, and could thereby acquire the skills needed to solve entirely new problems.

Although Lake (2019) argued that meta-learning "has implications for understanding how people generalize compositionally," he did not conduct a direct comparison to human behavior. In a follow-up study, Lake and Baroni (2023) addressed this shortcoming and found that meta-learned models "mimic human systematic generalization in a head-to-head comparison." These results are further corroborated by a recent paper of Jagadish, Binz, Saanum, Wang, and Schulz (2023) which demonstrated that meta-learned models capture human zero-shot compositional inferences in a reinforcement learning setting. However, there also remain open challenges in this context. For example, meta-learned models do not always generalize systematically to longer sequences than those in the training data (Lake, 2019; Lake & Baroni, 2023). How to resolve this issue will be an important challenge for future work.

3.3 Inductive biases

Human cognition comes with many useful inductive biases beyond the ability to reason compositionally. The preference for simplicity is one of these biases (Chater & Vitányi, 2003; Feldman, 2016). We readily extract abstract low-dimensional rules that allow us to generalize entirely new situations. Meta-learning is an ideal tool to build models with similar preferences because we can easily generate tasks based on simple rules and use them for meta-learning, thereby enabling an agent to acquire the desired inductive bias from data.

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

Toward this end, Kumar, Dasgupta, Cohen, Daw, and Griffiths (2020b) tested humans and meta-reinforcement agents on a grid-based task. People, as well as agents, encountered a series of 7 × 7 grids. Initially, all tiles were white, but clicking on them revealed their identity as either red or blue. The goal was to reveal all the red tiles while revealing as few blue tiles as possible. There was an underlying pattern that determined how the red tiles were placed, which was either specified by a structured grammar or by a nonstructured process with matched statistics. Humans found it easier to learn in structured tasks, confirming that they have strong priors toward simple abstract rules (Schulz, Tenenbaum, Duvenaud, Speekenbrink, & Gershman, 2017). However, their analysis also indicated that meta-learning is easier on nonstructured tasks than on structured tasks. In follow-up work, they found that this result also holds for agents that were trained purely on the structured version of their task but evaluated on both versions (Kumar et al., 2022a) – a quite astonishing finding considering that one would expect an agent to perform better on the task distribution it was trained on. The authors addressed this mismatch between humans and meta-learned agents by guiding agents during training to reproduce natural language descriptions that people provided to describe a given task. They found that grounding meta-learned agents in natural language descriptions not only improved their performance but also led to more human-like inductive biases, demonstrating that natural language can serve as a source for abstractions within human cognition.

Their line of work uses another interesting technique for training meta-learning agents (Kumar et al., 2022a, 2022b). It does not rely on a hand-designed task distribution but instead involves sampling tasks from the prior distribution of human participants using a technique known as Gibbs sampling with people (Harrison et al., 2020; Sanborn & Griffiths, 2007). Although doing so provides them with a data-set of tasks, no expression of the corresponding prior distribution over them is accessible and, hence, it is nontrivial to define a Bayesian model for the given setting. A meta-learned agent, on the other hand, was readily obtained by training on the collected samples.

3.4 Model-based reasoning

Many realistic scenarios afford two distinct types of learning: model-free and model-based. Model-free learning algorithms directly adjust their strategies using observed outcomes. Model-based learning algorithms, on the other hand, learn about the transition and reward probabilities of an environment, which are then used for downstream reasoning tasks. People are generally thought to be able to perform model-based learning, at least to some extent, and assuming that the problem at hand calls for it (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Kool, Cushman, & Gershman, 2016). Wang et al. (2016) showed that a meta-learned algorithm can display model-based behavior, even if it was trained through a pure model-free reinforcement learning algorithm (see Fig. 4c).

Having a model of the world also acts as the basis for causal reasoning. Traditionally, making causal inferences relies on the notion of Pearl's do-calculus (Pearl, 2009). Dasgupta et al. (2019), however, showed that meta-learning can be used to create models that draw causal inferences from observational data, select informative interventions, and make counterfactual predictions. Although they have not related their model to human data directly, it could in future work serve as the basis to study how people make causal judgments in complex domains and explain why and when they deviate from normative causal theories (Bramley, Dayan, Griffiths, & Lagnado, 2017; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021).

Together, these two examples highlight that model-based reasoning capabilities can emerge internally in a meta-learned model if they are beneficial for solving the encountered problem. Although there are already many traditional models that can perform such tasks, these models are often slow at run-time as they typically involve Bayesian inference, planning, or both. Meta-learning, on the other hand, "shifts most of the compute burden from inference time to training time [which] is advantageous when training time is ample but fast answers are needed at run-time" (Dasgupta et al., 2019), and may therefore explain how people can perform such intricate computations within a reasonable time frame.

Although model-based reasoning is an emerging property of meta-learned models, it may also be integrated explicitly into such models should it be desired. Jensen, Hennequin, and Mattar (2023) have taken this route, and augmented a standard meta-reinforcement learning agent with the ability to perform temporally extended planning using imagined rollouts. In each time-step, their agent can decide to perform a planning operation instead of directly interacting with the environment (in this case, a spatial navigation task). Their meta-learned agents opted to perform this planning operation consistently after training. Importantly, the model showed striking similarities to patterns of human deliberation by performing more planning early on and with an increased distance to the goal. Furthermore, they found that patterns of hippocampal replays resembled the rollouts of their model.

3.5 Exploration

People do not only have to integrate observed information into their existing knowledge, but they also have to actively determine what information to sample. They constantly face situations that require them to decide whether they should explore something new or whether they should rather exploit what they already know. Previous research suggests that people solve this exploration–exploitation dilemma using a combination of directed and random exploration strategies (Gershman, 2018; Schulz & Gershman, 2019; Wilson, Geana, White, Ludvig, & Cohen, 2014; Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018). Why do people use these particular strategies and not others? Binz and Schulz (2022a) hypothesized that they do so because human exploration follows resource-rational principles. To test this claim, they devised a family of resource-rational reinforcement learning algorithms by combining ideas from meta-learning and information theory. Their meta-learned model discovered a diverse set of exploration strategies, including random and directed exploration, that captured human exploration better than alternative approaches. In this domain, meta-learning offered a direct path toward investigating the hypothesis that people try to explore optimally but are subject to limited computational resources, whereas designing hand-crafted models for studying the same question would have been more intricate.

It is not only important to decide how to explore, but also to decide whether exploration is worthwhile in the first place. Lange and Sprekeler (2020) studied this question using the meta-learning framework. Their meta-learned agents are able to flexibly interpolate between implementing exploratory learning

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

behaviors and hard-coded, nonlearning strategies. Importantly, which behavior was realized crucially depended on environmental properties, such as the diversity of the task distribution, the task complexity, and the agent's lifetime. They showed, for instance, that agents with a short lifetime should opt for small rewards that are easy to find, whereas agents with an extended lifetime should spend their time exploring the environment. The study of Lange and Sprekeler (2020) clearly demonstrates that meta-learning makes it conceptually easy to iterate over different environmental assumptions inside a rational analysis of cognition. They only had to modify the environment as desired, followed by rerunning their meta-learning procedure. In contrast, traditional modeling approaches would require hand-designing a new optimal agent each time an environmental change occurs.

3.6 Cognitive control

Humans are remarkable at adapting to task-specific demands. The processes behind this ability are collectively referred to as cognitive control (Botvinick, Braver, Barch, Carter, & Cohen, 2001). Cohen (2017) even argues that "the capacity for cognitive control is perhaps the most distinguishing characteristic of human behavior." It should therefore come as no surprise that cognitive control has received a significant amount of attention from a computational perspective (Botvinick & Cohen, 2014; Collins & Frank, 2013). Recently, some of these computational investigations have been extended to the meta-learning framework.

The ability to adjust computational resources as needed is one hallmark of cognitive control. Moskovitz, Miller, Sahani, and Botvinick (2022) proposed a meta-learned model with such characteristics. Their model learns a simple default policy – similar to the model of Binz and Schulz (2022a) – that can be overwritten by a more complex one if necessary. They demonstrate that this model is not only able to capture behavioral phenomena from the cognitive control literature but also known effects in decision-making and reinforcement learning tasks, thereby linking the three domains. Importantly, their study highlights that the meta-learning framework offers the means to account for multiple computational costs instead of just a single one – in this case, a cost for implementing the default policy and one for deviating from it.

Taking contextual cues into consideration is another vital aspect of cognitive control. Dubey, Grant, Luo, Narasimhan, and Griffiths (2020) implemented this idea in the meta-learning framework. In their model, contextual cues determine the initialization of a task-specific neural network that is then trained using model-agnostic meta-learning. They showed that such a model captures "the context-sensitivity of human behavior in a simple but well-studied cognitive control task." Furthermore, they demonstrated that it scales well to more complex domains (including tasks from the MuJoCo [Todorov, Erez, & Tassa, 2012], CelebA [Liu, Luo, Wang, & Tang, 2015], and MetaWorld [Yu et al., 2020] benchmarks), thereby opening up new opportunities for modeling human behavior in naturalistic scenarios.

4. Why is not everything meta-learned?

We have laid out different arguments that make meta-learning a useful tool for constructing cognitive models, but it is important to note that we do not claim that meta-learning is the ultimate solution to every modeling problem. Instead, it is essential to understand when meta-learning is the right tool for the job and when not.

4.1 Lack of interpretability

Perhaps its most significant detriment is that meta-learning never provides us with analytical solutions that we can inspect, analyze, and reason about. In contrast to this, some Bayesian models have analytical solutions. Take as an example the data-generating distribution that we encountered earlier (Equations (1)–(2)). For these assumptions, a closed-form expression of the posterior predictive distribution is available. By looking at this closed-form expression, researchers have generated new predictions and subsequently tested them empirically (Daw et al., 2008; Dayan & Kakade, 2000; Gershman, 2015). Performing the same kind of analysis with a meta-learned model is not as straightforward. We do not have access to an underlying mathematical expression, which makes a structured exploration of theories much harder.

That being said, there are still ways to analyze a meta-learned model's behavior. For one, it is possible to use model architectures that facilitate interpretability. Binz et al. (2022) relied on this approach and designed a neural network architecture that produced weights of a probit regression model that were then used to cluster applied strategies into different categories. Doing so enabled them to identify which strategy was used by their meta-learned model in a particular situation.

Recently, researchers have also started to use tools from cognitive psychology to analyze the behavior of black-box models (Bowers et al., 2022; Rich & Gureckis, 2019; Ritter, Barrett, Santoro, & Botvinick, 2017; Schulz & Dayan, 2020). For example, it is possible to treat such models just like participants in a psychological experiment and use the collected data to analyze their behavior similar to how psychologists would analyze human behavior (Binz & Schulz, 2023; Dasgupta et al., 2022; Rahwan et al., 2019; Schramowski, Turan, Andersen, Rothkopf, & Kersting, 2022). We believe that this approach has great potential for analyzing increasingly capable and opaque artificial agents, including those obtained via meta-learning.

4.2 Intricate training processes

When using the meta-learning framework, one also has to deal with the fact that training neural networks is complex and takes time. Neural network models contain many moving parts, like weight initializations or the used optimizer, that have to be chosen appropriately such that training can take off in the first place, and training itself may take hours or days until it is finished. When we want to modify assumptions in the data-generating distribution, we have to retrain the whole system from scratch altogether. Thus, although the process of iterating over different environmental assumptions is conceptually straightforward in the meta-learning framework, it may be time consuming. Bayesian models can, in comparison, sometimes be more quickly adapted to changes in environmental assumptions. To illustrate this, let us assume that you wanted to explain human behavior through a meta-learned model that was trained on the data-generating distribution from Equations (1)–(2), but found that the resulting model does not fit the observed data well. Next, you want to consider the alternative hypothesis that people assume a nonstationary environment. Although this modification could be done quickly in the corresponding Bayesian model, the

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

meta-learning framework requires retraining on newly generated data.

There is, furthermore, no guarantee that a fully converged meta-learned model implements a Bayes-optimal learning algorithm. Indeed, there are reported cases in which meta-learning failed to find the Bayes-optimal solution (Wang et al., 2021). In simple scenarios, like our insect length example, we can resolve this issue by comparing to analytical solutions. This kind of reasoning applies to some of the settings in which meta-learning has been used to study human behavior. For example, for the exploration studies discussed in the previous section, it has been shown that meta-learned models closely approximate the (tractable but computationally expensive) Bayes-optimal algorithm (Duan et al., 2016; Wang et al., 2016). However, in more complex scenarios, it is impossible to verify that a meta-learned algorithm is optimal. We believe that this issue can be somewhat mitigated by validating meta-learned models in various ways. For example, we may get an intuition for the correspondence between a meta-learned model and an intractable Bayes-optimal algorithm by comparing to other approximate inference techniques (as done in Binz et al., 2022) or to symbolic models (as done in Lake & Baroni, 2023). In the end, however, we believe that this issue is still an open problem and that future work needs to come up with novel techniques to verify meta-learned models. Nevertheless, this is already a step forward as verifying solutions is often easier than generating them.

4.3 Meta-learned or Bayesian inference?

In summary, both frameworks – meta-learning and Bayesian inference – have their unique strengths and weaknesses. The meta-learning framework does and will not replace Bayesian inference but complement it. It broadens our available toolkit and enables researchers to study questions that were previously out of reach. However, there are certainly situations in which traditional Bayesian inference is a more appropriate modeling choice as we have outlined in this section.

5. The role of neural networks

Most of the points we have discussed so far are agnostic regarding the function approximator implementing the meta-learned algorithm. However, at the same time, we have appealed to neural networks at various points throughout the text. When one looks at prior work, it can also be observed that neural networks are the predominant model class in the meta-learning setting. Why is that the case? In addition to their universality, neural networks offer one big opportunity: They provide a flexible framework for engineering different types of inductive biases into a computational model (Goyal & Bengio, 2022). In the following section, we will highlight three examples of how previous work has accomplished this. For each of these examples, we take a concept from psychology, and show how it can be readily accommodated in a meta-learned model.

Perhaps one of the most persuasive ideas in cognitive modeling is that of gradient-based learning. It is not only at the heart of one of the most influential models – the Rescorla–Wagner model (Gershman, 2015; Rescorla, 1972) – but also features prominently in many other theories of human learning, such as connectionist models (Rumelhart et al., 1988). Even though the earlier outlined meta-learning procedure relies on gradient-based learning in the outer-loop, the resulting inner-loop dynamics must bear no resemblance to gradient descent. However, it is possible to construct meta-learned models whose inner-loop updates rely on gradient-based learning. Finn et al. (2017) proposed a meta-learning technique known as model-agnostic meta-learning that finds optimal initial parameters of a feedforward neural network that is subsequently trained via gradient descent. The idea is that these optimal initial parameters allow the feedforward network to generalize to multiple tasks in a minimal number of gradient steps. Although their general setup is similar to the one we discussed, it leads to models that learn via gradient descent instead of models that implement a learning algorithm inside the dynamics of a recurrent neural network. Kirsch and Schmidhuber (2021) recently brought these two approaches together into a single model. Their proposed architecture consists of multiple recurrent neural networks that interact with each other. Importantly, they showed that one particular configuration of these networks could implement backpropagation in the forward pass, thereby being able to perform gradient-based learning in a memory-based system.

Exemplar-based models – like the generalized category model (Nosofsky, 2011) – are one of the most prominent approaches for modeling how people categorize items into different classes (Kruschke, 1990; Shepard, 1987). They categorize a new instance based on the estimated similarity between that instance and previously seen examples. Recently, meta-learned models with exemplar-based reasoning abilities have been proposed for the task of few-shot classification, in which a classifier must generalize based on a training set containing only a few examples. Matching networks (Vinyals et al., 2016) accomplish this by classifying a new data-point using a similarity-weighted combination of categories in the training set. Importantly, similarity is computed over a learned embedding space, thereby ensuring that the model can scale to high-dimensional stimuli. Follow-up work has taken inspiration from another hugely influential model of human category learning and replaced the exemplar-based mechanism used in matching networks with one based on category prototypes (Snell, Swersky, & Zemel, 2017).

Finally, making inferences using similarities to previous experiences is not only useful for supervised learning but also in the reinforcement learning setting. In the reinforcement learning literature, the ability to store and recollect states or trajectories for later use is studied under the name of episodic memory (Lengyel & Dayan, 2007). It has been argued that episodic memory could be the key to explaining human performance in naturalistic environments (Gershman & Daw, 2017). Episodic memory also plays a crucial role in neuroscience, with studies showing that highly rewarding instances are stored in the hippocampus and made available for recall as and when required (Blundell et al., 2016). Ritter et al. (2018) build upon the neural episodic control idea from Pritzel et al. (2017) and use a differential neural dictionary for episodic recall in the context of meta-learning. Their dictionary stores encodings from previously experienced tasks, which can then be later queried as needed. With this addition, their meta-learned model is able to recall previously discovered policies, retrieve memories using category examples, handle compositional tasks, re-instate memories while traversing the environment, and recover a learning strategy people use in a neuroscience-inspired task.

In summary, human cognition comes with a variety of inductive biases and neural networks provide flexible ways to easily incorporate them into meta-learned models of cognition. We have outlined three such examples in the section, demonstrating

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

how to integrate gradient-based learning, exemplar- and prototype-based reasoning, and episodic memory into a meta-learned model. There are, furthermore, many other inductive biases for neural network architectures that could be used in the context of meta-learning but have not been yet. Examples include networks that perform differentiable planning (Farquhar, Rocktäschel, Igl, & Whiteson, 2017; Tamar, Wu, Thomas, Levine, & Abbeel, 2016), extract object-based representations (Piloto, Weinstein, Battaglia, & Botvinick, 2022; Sancaktar, Blaes, & Martius, 2022), or modify their own connections through synaptic plasticity (Miconi, Rawal, Clune, & Stanley, 2020; Schlag, Irie, & Schmidhuber, 2021).

6. Toward a domain-general model of human learning

What does the future hold for meta-learning? The current generation of meta-learned models of cognition is almost exclusively trained on the data-generating distribution of a specific problem family. Although this training process enables them to generalize to new tasks inside this problem family, they are unlikely to generalize to completely different domains. We would, for example, not expect a meta-learned algorithm to perform a challenging maze navigation task if it was only trained to predict the lengths of insect species.

Although domain-specific models have (and will continue to) provide answers to important research questions, we agree with Newell (1992) that "unified theories of cognition are the only way to bring this wonderful, increasing fund of knowledge under intellectual control." Ideally, such a unified theory should manifest itself in a domain-general cognitive model that cannot only solve prediction tasks but is also capable of human-like decision making (Gigerenzer & Gaissmaier, 2011), category learning (Ashby et al., 2005), navigation (Montello, 2005), problem-solving (Newell et al., 1972), and so forth. We consider the meta-learning framework the ideal tool for accomplishing this goal as it allows us to compile arbitrary assumptions about an agent's beliefs of the world (arguments 1 and 2) and its computational architecture (arguments 3 and 4) into a cognitive model.

To obtain such a domain-general cognitive model via meta-learning, however, a few challenges need to be tackled. First of all, there is the looming question of how a data-generating distribution that contains many different problems should be constructed. Here, we may take inspiration from the machine learning community, where researchers have devised generalist agents by training neural networks on a large set of problems (Reed et al., 2022). Team et al. (2023) have recently shown that this is a promising path for scaling up meta-learning models. They trained a meta-reinforcement learning agent on a vast open-ended world with over 10^{40} possible tasks. The resulting agent can adapt to held-out problems as quickly as humans, and "displays on-the-fly hypothesis-driven exploration, efficient exploitation of acquired knowledge, and can successfully be prompted with first-person demonstrations." In the same vein, we may come up with a large collection of tasks that are more commonly used to study human behavior (Miconi, 2023; Molano-Mazon et al., 2022; Yang et al., 2019), and use them to train a meta-learned model of cognition.

Language will likely play an important role in future meta-learning systems. We do not want a system that learns every task from scratch via trial and error but one that can be provided with a set of instructions similar to how a human subject would be instructed in a psychological experiment. Having agents capable of language will not only enable them to understand new tasks in a zero-shot manner but may also facilitate their cognitive abilities. It, for example, allows them to decompose tasks into subtasks, learn from other agents, or generate explanations (Colas, Karch, Moulin-Frier, & Oudeyer, 2022). Fortunately, our current best language models (Brown et al., 2020; Chowdhery et al., 2022) and meta-learning systems are both based on neural networks. Thus, integrating language capabilities into a meta-learned model of cognition should – at least conceptually – be fairly straightforward. Doing so would furthermore enable such models to harvest the compositional nature of language to make strong generalizations to tasks outside of the meta-learning distribution. The potential for this was highlighted in a recent study (Riveland & Pouget, 2022) which found that language-conditioned recurrent neural network models can perform entirely novel psychophysical tasks with high accuracy.

Moreover, a sufficiently general model of human cognition must not only be able to select among several given options, like in a decision-making or category learning setting, but it also needs to maneuver within a complex world. For this, it needs to perceive and process high-dimensional visual stimuli, it needs to control a body with many degrees of freedom, and it needs to actively engage with other agents. Many of these problems have been at the heart of the deep learning community (Hill et al., 2020; McClelland, Hill, Rudolph, Baldridge, & Schütze, 2020; Strouse, McKee, Botvinick, Hughes, & Everett, 2021; Team et al., 2021), and it will be interesting to see whether the solutions developed there can be integrated into a meta-learned model of cognition.

Finally, there are also some challenges on the algorithmic side that need to be taken into account. In particular, it is unclear how far currently used model architectures and outer-loop learning algorithms scale. Although contemporary meta-learning algorithms are able to find approximately Bayes-optimal solutions to simple problems, they sometimes struggle to do so on more complex ones (e.g., as in the earlier discussed work of Wang et al., 2021). Therefore, it seems likely that simply increasing the complexity of the meta-learning distribution will not be sufficient – we will also need model architectures and outer-loop learning algorithms that can handle increasingly complex data-generating distributions. The transformer architecture (Vaswani et al., 2017), which has been very successful at training large language models (Brown et al., 2020; Chowdhery et al., 2022), provides one promising candidate, but there could be countless other (so far undiscovered) alternatives.

Thus, taken together, there are still substantial challenges involved in creating a domain-general meta-learned model of cognition. In particular, we have argued in this section that we need to (1) meta-learn on more diverse task distributions, (2) develop agents that can parse instructions in the form of natural language, (3) embody these agents in realistic problem settings, and (4) find model architectures that scale to these complex problems. Figure 5 summarizes these points graphically.

7. Conclusion

Most computational models of human learning are hand-designed, meaning that at some point a researcher sat down and defined how they behave. Meta-learning starts with an entirely different premise. Instead of designing learning algorithms by hand, one trains a system to achieve its goals by repeatedly letting it interact with an environment. Although this seems

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

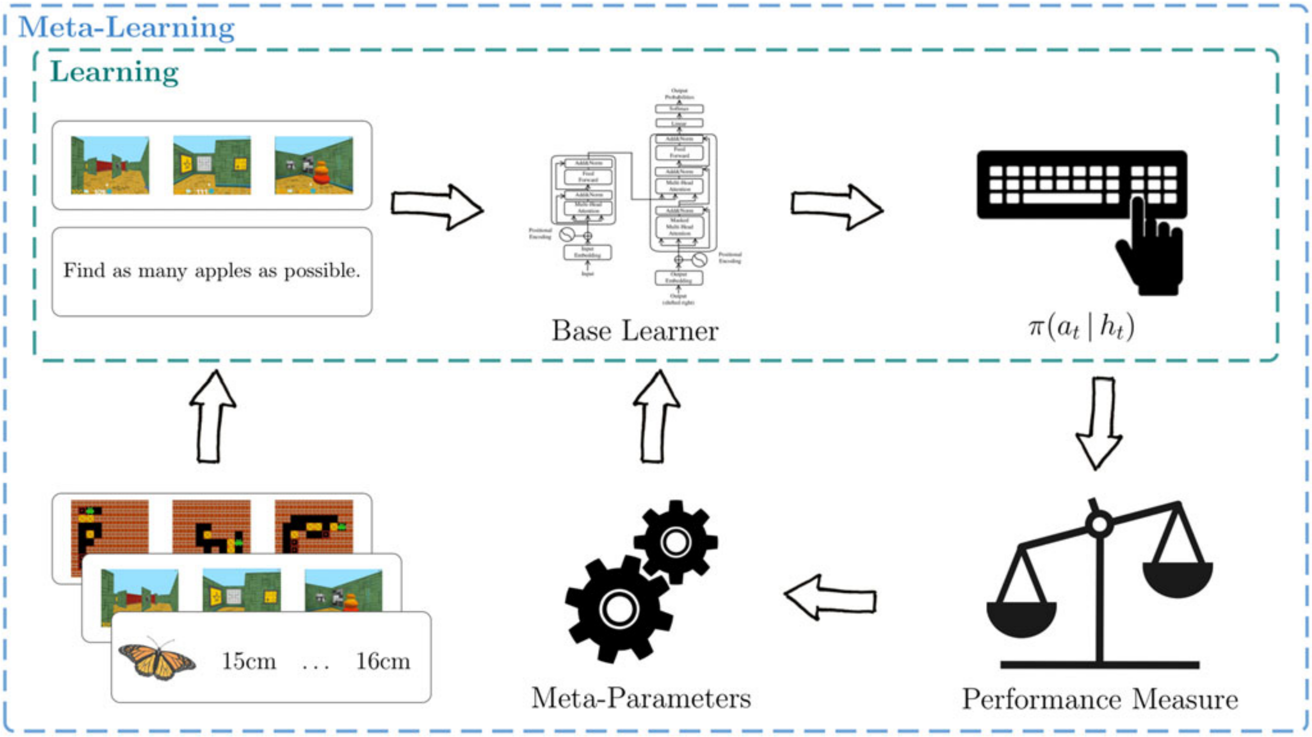


Figure 5. Illustration of how a domain-general meta-learned model of cognition could look like. Modifications include training on more diverse task distributions, providing natural language instructions as additional inputs, and relying on scalable model architectures.

quite different from traditional models of learning on the surface, we have highlighted that the meta-learning framework actually has a deep connection to the idea of Bayesian inference, and thereby to the rational analysis of cognition. Using this connection as a starting point, we have highlighted several advantages of the meta-learning framework for constructing rational models of cognition. Together, our arguments demonstrate that meta-learning cannot only be applied in situations where Bayesian inference is impossible but also facilitates the inclusion of computational constraints and neuroscientific insights into rational models of human cognition. Earlier criticisms of the rational analysis of cognition have repeatedly pointed out that "rational Bayesian models are significantly unconstrained" and that they should be "developed in conjunction with mechanistic considerations to offer substantive explanations of cognition" (Jones & Love, 2011). Likewise, Bowers and Davis (2012) argued that to understand human cognition "important constraints [must] come from biological, evolutionary, and processing (algorithmic) considerations." We believe that the meta-learning framework provides the ideal opportunity to address these criticisms as it allows for a painless integration of flexible algorithmic (often biologically inspired) mechanisms.

It is worth pointing out that meta-learning can be also motivated by taking neural networks as a starting point. From this perspective, it bridges two of the most popular theories of cognition – Bayesian models and connectionism – by bringing the scalability of neural network models into the rational analysis of cognition. The blending of Bayesian models and neural networks situates the meta-learning framework at the heart of the debate on whether cognition is best explained by emergentist or structured probabilistic approaches (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; McClelland et al., 2010). Like traditional connectionist approaches, meta-learning provides a means to explain how cognition could emerge as a system repeatedly interacts with an environment. Whether the current techniques used for meta-learning mirror the emergence of cognitive processes in people however remains an open question. Personally, we believe that this is unlikely and that there are more elaborate processes in play during human meta-learning than the gradient descent-based algorithms that are commonly used for training neural networks (Schulze Buschoff, Schulz, & Binz, 2023). To study this question systemically, we would need to look at human behavior across much longer timescales (e.g., developmental or evolutionary). Yet, at the same time, meta-learning does not limit itself to a purely emergentist perspective. The modern neural network toolbox allows researchers to flexibly integrate additional structure and inductive biases into a model by adjusting the underlying network architecture – as we have argued in Section 5 – thereby preserving a key advantage of structured probabilistic approaches. How much hand-crafting within the network architecture is needed ultimately depends on the designer's goals. The meta-learning framework is agnostic to this and allows it to range from almost nothing to a substantial amount.

We believe that meta-learning provides a powerful tool to scale up psychological theories to more complex settings. However, at the same time, meta-learning has not delivered on this promise yet. Existing meta-learned models of cognition are typically applied to classical scenarios where established models already exist. Thus, we have to ask: What prevents the application to more complex and general paradigms? First, such paradigms themselves have to be developed. Fortunately, there is currently a trend toward measuring human behavior on more naturalistic tasks (Brändle, Binz, & Schulz, 2022a; Brändle, Stocks,

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

Tenenbaum, Gershman, & Schulz, 2022b; Schulz et al., 2019), and it will be interesting to see what role meta-learning will play in modeling behavior in such settings. Furthermore, meta-learning can be intricate and time consuming. We hope that the present article – together with the accompanying code examples – makes this technique less opaque and more accessible to a wider audience. Future advances in hardware will likely make the meta-learning process quicker and we are therefore hopeful that meta-learning can ultimately fulfill its promise of identifying plausible models of human cognition in situations that are out of reach for hand-designed algorithms.

Acknowledgments. The authors thank Sreejan Kumar, Tobias Ludwig, Dominik Endres, and Adam Santoro for their valuable feedback on an earlier draft.

Financial support. This work was funded by the Max Planck Society, the Volkswagen Foundation, as well as the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC2064/1-390727645.

Competing interest. None.

Notes

[1] Based on our earlier definition, it is at this point strictly speaking not a learning algorithm at all as it does not improve with additional data.

[2] https://github.com/marcelbinz/meta-learned-models.

[3] There has been a long-standing conceptual debate in cognitive psychology on whether to view Bayesian models as normative standards or descriptive tools. We believe that this debate is beyond the scope of the current article and thus refer the reader to earlier work for an in-depth discussion (Bowers & Davis, 2012; Griffiths, Chater, Norris, & Pouget, 2012; Jones & Love, 2011; Tauber, Navarro, Perfors, & Steyvers, 2017; Zednik & Jäkel, 2016). We only want to add that the framework outlined here is agnostic to this issue – meta-learned models may serve as both normative standards and descriptive tools.

[4] In principle, one could select arbitrarily flexible functional forms, such as mixtures of normal distributions or discretized distributions with small bin sizes, which would reduce the accompanying approximation error.

[5] This only holds for standard variational inference but not for more advanced methods that involve amortization such as variational autoencoders (Kingma & Welling, 2013).

[6] Note that although it is possible to apply some Bayesian models (e.g., non-parametric methods) in this setting, we would have to contend with making arbitrary assumptions about the likelihood function, causing a loss of optimality guarantees.

[7] Having said that, it is possible to approximate it under certain circumstances and different authors have applied such approximations to study both human and animal cognition (Chater & Vitányi, 2003; Gauvrit, Zenil, Delahaye, & Soler-Toscano, 2014; Gauvrit, Zenil, & Tegnér, 2017; Griffiths, Daniels, Austerweil, & Tenenbaum, 2018; Zenil, Marshall, & Tegnér, 2015).

References

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, *62*(3), 547–554.

Anderson, J. R. (2013a). *The adaptive character of thought*. Psychology Press.

Anderson, J. R. (2013b). *The architecture of cognition*. Psychology Press.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*(1), 149–178.

Bates, C. J., & Jacobs, R. A. (2020). Efficient data compression in perception and perceptual memory. *Psychological Review*, *127*(5), 891.

Baxter, J. (1998). Theoretical models of learning to learn. In S. Thrun & L. Pratt (Eds.), *Learning to learn* (pp. 71–94). Springer.

Bellec, G., Salaj, D., Subramoney, A., Legenstein, R., & Maass, W. (2018). Long short-term memory and learning-to-learn in networks of spiking neurons. *Advances in Neural Information Processing Systems*, *31*, 795–805.

Bengio, Y., Bengio, S., & Cloutier, J. (1991). Learning a synaptic learning rule. In *IJCNN-91-Seattle International Joint Conference on Neural Networks, Seattle, WA, USA* (Vol. 2, p. 969).

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. In B. Bernheim, S. DellaVigna, & D. Laibson (Eds.), *Handbook of behavioral economics: Applications and foundations* (Vol. 2, pp. 69–186). North-Holland.

Binmore, K. (2007). Rational decisions in large worlds. *Annales d'Economie et de Statistique*, *No. 86*, 25–41.

Binz, M., Gershman, S. J., Schulz, E., & Endres, D. (2022). Heuristics from bounded meta-learned inference. *Psychological Review*, *129*(5), 1042–1077.

Binz, M., & Schulz, E. (2022a). Modeling human exploration through resource-rational reinforcement learning. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in neural information processing systems* (pp. 31755–31768). Curran Associates, Inc. https://openreview.net/forum?id=W1MUJv5zaXP.

Binz, M., & Schulz, E. (2022b). Reconstructing the Einstellung effect. *Computational Brain & Behavior*, *6*, 1–17.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(6), e2218523120.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blundell, C., Uria, B., Pritzel, A., Li, Y., Ruderman, A., Leibo, J. Z., … Hassabis, D. (2016). Model-free episodic control. *arXiv preprint arXiv:1606.04460*.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624.

Botvinick, M. M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science*, *38*(6), 1249–1285.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*(3), 389.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., … Blything, R. (2022). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, 1–74.

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301.

Brändle, F., Binz, M., & Schulz, E. (2022a). Exploration beyond bandits. In I. Cogliati Dezza, E. Schulz, & C. M. Wu (Eds.), *The drive for knowledge: The science of human information seeking* (pp. 147–168). Cambridge University Press. doi:10.1017/9781009026949.008

Brändle, F., Stocks, L. J., Tenenbaum, J. B., Gershman, S. J., & Schulz, E. (2022b). Intrinsically motivated exploration as empowerment. *PsyArXiv*. January 14.

Brighton, H., & Gigerenzer, G. (2012). Are rational actor models "rational" outside small worlds. In S. Okasha & K. Binmore (Eds.), *Evolution and rationality: Decisions, co-operation, and strategic behavior* (pp. 84–109). Cambridge University Press.

Bromberg-Martin, E. S., Matsumoto, M., Hong, S., & Hikosaka, O. (2010). A pallidus–habenula–dopamine pathway signals inferred stimulus values. *Journal of Neurophysiology*, *104*(2), 1068–1076.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Chaitin, G. J. (1969). On the simplicity and speed of programs for computing infinite sets of natural numbers. *Journal of the ACM (JACM)*, *16*(3), 407–422.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*(2), 57–65.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22.

Chomsky, N. (2014). *Aspects of the theory of syntax* (Vol. 11). MIT Press.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., … Fiedel, N. (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Cohen, J. D. (2017). Cognitive control: Core constructs and current considerations. In T. Egner (Ed.), *The Wiley handbook of cognitive control* (pp. 1–28). Wiley Blackwell.

Colas, C., Karch, T., Moulin-Frier, C., & Oudeyer, P.-Y. (2022). Language and culture internalization for human-like autotelic AI. *Nature Machine Intelligence*, *4*(12), 1068–1076.

Collins, A. G., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review*, *120*(1), 190.

Corner, A., & Hahn, U. (2013). Normative theories of argumentation: Are some norms better than others? *Synthese*, *190*(16), 3579–3610.

Courville, A. C., & Daw, N. D. (2008). The rat as particle filter. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems* (pp. 369–376). Curran Associates, Inc.

Cover, T. M. (1999). *Elements of information theory*. John Wiley.

Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(48), 30055–30062.

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics. In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 97–118). Oxford University Press.

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

Dasgupta, I., & Gershman, S. J. (2021). Memory as a computational resource. *Trends in Cognitive Sciences*, *25*(3), 240–251.

Dasgupta, I., Lampinen, A. K., Chan, S. C., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2022). Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive Psychology*, *96*, 1–25.

Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review*, *127*(3), 412.

Dasgupta, I., Wang, J., Chiappa, S., Mitrovic, J., Ortega, P., Raposo, D., … Kurth-Nelson, Z. (2019). Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*.

Daw, N. D., Courville, A. C., & Dayan, P. (2008). Semi-rational models of conditioning: The case of trial order. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind* (pp. 431–452). Oxford Academic.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Dayan, P., & Kakade, S. (2000). Explaining away in weight space. *Advances in Neural Information Processing Systems*, *13*, 430–436.

Dobs, K., Martinez, J., Kell, A. J., & Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. *Science Advances*, *8*(11), eabl8913.

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*(4–6), 495–506.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., & Abbeel, P. (2016). RL2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*.

Dubey, R., Grant, E., Luo, M., Narasimhan, K., & Griffiths, T. (2020). Connecting context-specific adaptation in humans to meta-learning. *arXiv preprint arXiv:2011.13782*.

Duff, M. O. (2003). *Optimal learning: Computational procedures for Bayes-adaptive Markov decision processes* [Unpublished PhD thesis]. University of Massachusetts Amherst.

Farquhar, G., Rocktäschel, T., Igl, M., & Whiteson, S. (2017). TreeQN and ATreeC: Differentiable tree-structured models for deep reinforcement learning. *arXiv preprint arXiv:1710.11417*.

Feldman, J. (2016). The simplicity principle in perception and cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(5), 330–340.

Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), *Automated machine learning* (pp. 3–33). Springer.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning* (pp. 1126–1135).

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Gauvrit, N., Zenil, H., Delahaye, J.-P., & Soler-Toscano, F. (2014). Algorithmic complexity for short binary strings applied to psychology: A primer. *Behavior Research Methods*, *46*(3), 732–744.

Gauvrit, N., Zenil, H., & Tegnér, J. (2017). The information-theoretic and algorithmic approach to human, animal, and artificial cognition. In G. Dodig-Crnkovic & R. Giovagnoli (Eds.), *Representation and reality in humans, other living organisms and intelligent machines* (pp. 117–139). Springer.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6), 721–741.

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLoS Computational Biology*, *11*(11), e1004567.

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42.

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, *68*, 101.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(5), 936.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B: Methodological*, *41*(2), 148–177.

Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, *478*(2266), 20210068.

Grant, E., Finn, C., Levine, S., Darrell, T., & Griffiths, T. (2018). Recasting gradient-based meta-learning as hierarchical Bayes. In *6th international conference on learning representations, ICLR 2018*.

Griffiths, T. L., Callaway, F., Chang, M. B., Grant, E., Krueger, P. M., & Lieder, F. (2019). Doing more with less: Meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, *29*, 24–30.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin*, *138*(3), 415–422.

Griffiths, T. L., Daniels, D., Austerweil, J. L., & Tenenbaum, J. B. (2018). Subjective randomness as statistical inference. *Cognitive Psychology*, *103*, 85–109.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 59–100). Cambridge University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.

Harlow, H. F. (1949). The formation of learning sets. *Psychological Review*, *56*(1), 51.

Harrison, P., Marjieh, R., Adolfi, F., van Rijn, P., Anglada-Tort, M., Tchernichovski, O., … Jacoby, N. (2020). Gibbs sampling with people. *Advances in Neural Information Processing Systems*, *33*, 10659–10671.

Hill, F., Lampinen, A., Schneider, R., Clark, S., Botvinick, M., McClelland, J. L., & Santoro, A. (2020). Environmental drivers of systematicity and generalization in a situated agent. In *International conference on learning representations*. Retrieved from https://openreview.net/forum?id=SklGryBtwr

Hinton, G. E., & Van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the 6th annual conference on computational learning theory* (pp. 5–13).

Hinton, G. E., & Zemel, R. (1993). Autoencoders, minimum description length and Helmholtz free energy. *Advances in Neural Information Processing Systems*, *6*, 3–10.

Hochreiter, S., Younger, A. S., & Conwell, P. R. (2001). Learning to learn using gradient descent. In *International conference on artificial neural networks* (pp. 87–94).

Hoppe, D., & Rothkopf, C. A. (2016). Learning rational temporal eye movement strategies. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(29), 8332–8337.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.

Jagadish, A. K., Binz, M., Saanum, T., Wang, J. X., & Schulz, E. (2023). Zero-shot compositional reinforcement learning in humans. https://doi.org/10.31234/osf.io/ymve5.

Jensen, K. T., Hennequin, G., & Mattar, M. G. (2023). A recurrent network model of planning explains hippocampal replay and human behavior. *bioRxiv*, 2023-01.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(4), 169.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, *37*(2), 183–233.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237.

Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask "why" questions of minds and brains. *Trends in Neurosciences*, *46*(3), 240–254.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Kirsch, L., & Schmidhuber, J. (2021). Meta learning backpropagation and improving it. *Advances in Neural Information Processing Systems*, *34*, 14122–14134.

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge University Press.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, *1*(1), 1–7.

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When does model-based control payoff? *PLoS Computational Biology*, *12*(8), e1005090.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247.

Kruschke, J. (1990). Alcove: A connectionist model of human category learning. *Advances in Neural Information Processing Systems*, *3*, 649–655.

Kumar, S., Correa, C. G., Dasgupta, I., Marjieh, R., Hu, M., Hawkins, R. D., … Griffiths, T. L. (2022a). Using natural language and program abstractions to instill human inductive biases in machines. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in neural information processing systems* (pp. 167–180). Curran Associates, Inc. https://openreview.net/forum?id=buXZ7nIqiwE.

Kumar, S., Dasgupta, I., Cohen, J., Daw, N., & Griffiths, T. (2020b). Meta-learning of structured task distributions in humans and machines. In *International conference on learning representations*.

Kumar, S., Dasgupta, I., Cohen, J. D., Daw, N. D., & Griffiths, T. L. (2020a). Meta-learning of structured task distributions in humans and machines. *arXiv preprint arXiv:2010.02317*.

Kumar, S., Dasgupta, I., Marjieh, R., Daw, N. D., Cohen, J. D., & Griffiths, T. L. (2022b). Disentangling abstraction from statistical pattern matching in human and machine learning. *arXiv preprint arXiv:2204.01437*.

Lake, B. M. (2019). Compositional generalization through meta sequence-to-sequence learning. *Advances in Neural Information Processing Systems*, *32*, 9791–9801.

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

Lake, B. M., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, *623*, 1–7.

Lange, R. T., & Sprekeler, H. (2020). Learning not to learn: Nature versus nurture in silico. *arXiv preprint arXiv:2010.04466*.

Lengyel, M., & Dayan, P. (2007). Hippocampal contributions to control: The third way. *Advances in Neural Information Processing Systems*, *20*, 889–896.

Lewis, D. (1999). Why conditionalize? In Lewis D. (Ed.) , *Papers in metaphysics and epistemology* (Vol. 2, pp. 403–407). Cambridge University Press. doi:10.1017/CBO9780511625343.024

Li, Z., Zhou, F., Chen, F., & Li, H. (2017). Meta-SGD: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *124*(6), 762.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015, December). Deep learning face attributes in the wild. In *Proceedings of international conference on computer vision (ICCV)*.

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215.

Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P., & Macke, J. (2021). Benchmarking simulation-based inference. In *International conference on artificial intelligence and statistics* (pp. 343–351).

Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348–356.

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. (2020). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(42), 25966–25974.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419.

McCoy, R. T., Grant, E., Smolensky, P., Griffiths, T. L., & Linzen, T. (2020). Universal linguistic inductive biases via meta-learning. *arXiv preprint arXiv:2006.16324*.

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

Miconi, T. (2023). A large parametrized space of meta-reinforcement learning tasks. *arXiv preprint arXiv:2302.05583*.

Miconi, T., Rawal, A., Clune, J., & Stanley, K. O. (2020). Backpropamine: Training self-modifying neural networks with differentiable neuromodulated plasticity. *arXiv preprint arXiv:2002.10585*.

Mikulik, V., Delétang, G., McGrath, T., Genewein, T., Martic, M., Legg, S., & Ortega, P. (2020). Meta-trained agents implement Bayes-optimal agents. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 18691–18703). Curran. Retrieved from https://proceedings.neurips.cc/paper/2020/file/d902c3ce47124c66ce615d5ad9ba304f-Paper.pdf

Mitchell, T. M. (1997). *Machine learning* (Vol. 1). McGraw Hill.

Molano-Mazon, M., Barbosa, J., Pastor-Ciurana, J., Fradera, M., Zhang, R.-Y., Forest, J., … Yang, G. R. (2022). Neurogym: An open resource for developing and sharing neuroscience tasks. https://doi.org/10.31234/osf.io/aqc9n.

Montello, D. R. (2005). *Navigation*. Cambridge University Press.

Moskovitz, T., Miller, K., Sahani, M., & Botvinick, M. M. (2022). A unified theory of dual-process control. *arXiv preprint arXiv:2211.07036*.

Müller, S., Hollmann, N., Arango, S. P., Grabocka, J., & Hutter, F. (2021). Transformers can do Bayesian inference. *arXiv preprint arXiv:2112.10510*.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Newell, A. (1992). Unified theories of cognition and the role of soar. In *Soar: A cognitive architecture in perspective* (pp. 25–79). Springer.

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104, No. 9). Prentice Hall.

Nichol, A., Achiam, J., & Schulman, J. (2018). On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 18–39). Cambridge University Press.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Ortega, P. A., Braun, D. A., Dyer, J., Kim, K.-E., & Tishby, N. (2015). Information-theoretic bounded rationality. *arXiv preprint arXiv:1512.06789*.

Ortega, P. A., Wang, J. X., Rowland, M., Genewein, T., Kurth-Nelson, Z., Pascanu, R., … Legg, S. (2019). Meta-learning of sequential strategies. *arXiv preprint arXiv:1905.03030*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran. Retrieved from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pearl, J. (2009). *Causality*. Cambridge University Press.

Piloto, L. S., Weinstein, A., Battaglia, P., & Botvinick, M. (2022). Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, *6*(9), 1257–1267.

Pritzel, A., Uria, B., Srinivasan, S., Badia, A. P., Vinyals, O., Hassabis, D., … Blundell, C. (2017). Neural episodic control. In *International conference on machine learning* (pp. 2827–2836).

Rabinowitz, N. C. (2019). Meta-learners' learning dynamics are unlike learners'. *arXiv preprint arXiv:1905.01320*.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922.

Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., … de Freitas, N. (2022). A generalist agent. *arXiv preprint arXiv:2205.06175*.

Rescorla, M. (2020). An improved Dutch book theorem for conditionalization. *Erkenntnis*, *87*, 1–29.

Rescorla, R. A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Current research and theory* (pp. 64–99). Appleton-Century-Crofts.

Rich, A. S., & Gureckis, T. M. (2019). Lessons for artificial intelligence from the study of natural stupidity. *Nature Machine Intelligence*, *1*(4), 174–180.

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. In *International conference on machine learning* (pp. 2940–2949).

Ritter, S., Wang, J., Kurth-Nelson, Z., Jayakumar, S., Blundell, C., Pascanu, R., & Botvinick, M. (2018). Been there, done that: Meta-learning with episodic recall. In *International conference on machine learning* (pp. 4354–4363).

Riveland, R., & Pouget, A. (2022). Generalization in sensorimotor networks configured with natural language instructions. *bioRxiv*, 2022-02.

Rosenkrantz, R. D. (1992). The justification of induction. *Philosophy of Science*, *59*(4), 527–539.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1988). *Parallel distributed processing* (Vol. 1). IEEE.

Sanborn, A., & Griffiths, T. (2007). Markov chain Monte Carlo with people. *Advances in Neural Information Processing Systems*, *20*, 1265–1272.

Sanborn, A. N. (2017). Types of approximation for probabilistic cognition: Sampling and variational. *Brain and Cognition*, *112*, 98–101.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144.

Sanborn, A. N., & Silva, R. (2013). Constraining bridges between levels of analysis: A computational justification for locally Bayesian learning. *Journal of Mathematical Psychology*, *57*(3–4), 94–106.

Sancaktar, C., Blaes, S., & Martius, G. (2022). Curious exploration via structured world models yields zero-shot object manipulation. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.) , *Advances in neural information processing systems* (pp. 24170–24183). Curran Associates, Inc. https://openreview.net/forum?id=NnuYZ1el24C.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. (2016). Meta-learning with memory-augmented neural networks. In *International conference on machine learning* (pp. 1842–1850).

Savage, L. J. (1972). *The foundations of statistics*. Courier.

Schaul, T., & Schmidhuber, J. (2010). Metalearning. *Scholarpedia*, *5*(6), 4650 (revision #91489). doi:10.4249/scholarpedia.4650

Schlag, I., Irie, K., & Schmidhuber, J. (2021). Linear transformers are secretly fast weight programmers. In *International conference on machine learning* (pp. 9355–9366).

Schmidhuber, J. (1987). Evolutionary principles in self-referential learning, or on learning how to learn: The meta-meta-… hook. Unpublished doctoral dissertation, Technische Universität München.

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, *4*(3), 258–268.

Schulz, E., Bhui, R., Love, B. C., Brier, B., Todd, M. T., & Gershman, S. J. (2019). Structured, uncertainty-driven exploration in real-world consumer choice. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(28), 13903–13908.

Schulz, E., & Dayan, P. (2020). Computational psychiatry for computers. *iScience*, *23*(12), 101772.

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, *55*, 7–14.

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive Psychology*, *99*, 44–79. doi:10.1016/j.cogpsych.2017.11.002

Schulze Buschoff, L. M., Schulz, E., & Binz, M. (2023). The acquisition of physical knowledge in generative neural networks. In *Proceedings of the 40th international conference on machine learning* (pp. 30321–30341).

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, *30*, 4080–4090.

Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I. *Information and Control*, *7*(1), 1–22.

Stopper, C. M., Maric, T., Montes, D. R., Wiedman, C. R., & Floresco, S. B. (2014). Overriding phasic dopamine signals redirects action selection during risk/reward decision making. *Neuron*, *84*(1), 177–189.

Strouse, D., McKee, K., Botvinick, M., Hughes, E., & Everett, R. (2021). Collaborating with humans without human data. *Advances in Neural Information Processing Systems*, *34*, 14502–14515.

Tamar, A., Wu, Y., Thomas, G., Levine, S., & Abbeel, P. (2016). Value iteration networks. *Advances in Neural Information Processing Systems*, *29*, 2154–2162.

Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. *Psychological Review*, *124*(4), 410.

Team, A. A., Bauer, J., Baumli, K., Baveja, S., Behbahani, F., Bhoopchand, A., … Zhang, L. (2023). Human-timescale adaptation in an open-ended task space. *arXiv preprint arXiv:2301.07608*.

Team, O. E. L., Stooke, A., Mahajan, A., Barros, C., Deck, C., Bauer, J., … Czarnecki, W. M. (2021). Open-ended learning leads to generally capable agents. *arXiv preprint arXiv:2107.12808*.

Tenenbaum, J. (2021). Joshua Tenenbaum's homepage. Retrieved from http://web.mit.edu/cocosci/josh.html

Thrun, S., & Pratt, L. (1998). Learning to learn: Introduction and overview. In S. Thrun & L. Pratt (Eds.), *Learning to learn* (pp. 3–17). Springer.

Todorov, E., Erez, T., & Tassa, Y. (2012). Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems* (pp. 5026–5033).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 6000–6010.

Vinyals, O., Blundell, C., Lillicrap, T., & Wierstra, D. (2016). Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, *29*, 3637–3645.

Wang, J. X. (2021). Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences*, *38*, 90–95.

Wang, J. X., King, M., Porcel, N., Kurth-Nelson, Z., Zhu, T., Deck, C., … Botvinick, M. (2021). Alchemy: A structured task distribution for meta-reinforcement learning. CoRR, abs/2102.02926. Retrieved from https://arxiv.org/abs/2102.02926

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., … Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(6), 860–868.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., … Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, *8*(7), 1341–1390.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915–924.

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X.-J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, *22*(2), 297–306.

Yang, Y., & Piantadosi, S. T. (2022). One model for the learning of language. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(5), e2021865119.

Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., & Levine, S. (2020). Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning* (pp. 1094–1100).

Zednik, C., & Jäkel, F. (2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese*, *193*(12), 3951–3985.

Zenil, H., Marshall, J. A., & Tegnér, J. (2015). Approximations of algorithmic and structural complexity validate cognitive-behavioural experimental results. *arXiv preprint arXiv:1509.06338*.

Open Peer Commentary

Challenges of meta-learning and rational analysis in large worlds

Margherita Calderan[a] and Antonino Visalli[b]*

[a]Department of Developmental Psychology and Socialisation, University of Padova, Italy and [b]IRCCS San Camillo Hospital, Venice, Italy
margherita.calderan@phd.unipd.it
antonino.visalli@hsancamillo.it

*Corresponding author.

doi:10.1017/S0140525X24000128, e148

Abstract

We challenge Binz et al.'s claim of meta-learned model superiority over Bayesian inference for large world problems. While comparing Bayesian priors to model-training decisions, we question meta-learning feature exclusivity. We assert no special justification for rational Bayesian solutions to large world problems, advocating exploring diverse theoretical frameworks beyond rational analysis of cognition for research advancement.

Binz et al. (argument 2) advocate for the superiority of meta-learned models over Bayesian inference for addressing large world problems (Savage, 1972). Our commentary aims to question some perceived fallacies in their arguments.

First, although we recognize that "*Identifying the correct set of assumptions becomes especially challenging once we deal with more complex problems*," we point out that meta-learned models also require specific assumptions. Examples are the selection of samples from the data-generating distribution, choice of the optimizer, weight initializations, or constraints to mimic bounded rationality. These decisions, too, can be conceived as priors and require a certain level of justification. Binz et al. explicitly emphasized the importance of appropriately making these choices (sect. 4, "Intricate Training Processes"). Bayesian or not, prior knowledge is a necessary condition for both modeling procedures. As a consequence, we contend that both Bayesian and meta-learned models present similar challenges from a rational perspective. Therefore, why should it be "*hard to justify*" prior assumptions for Bayesian models and not for meta-learned models? For instance, one could reconsider the critiques moved to Lucas, Griffiths, Williams, and Kalish (2015). To account for the bias toward expecting linear relationships between continuous variables, the authors assigned lower prior probabilities to quadratic and radial relationships as compared to linear ones (Lucas et al., 2015). Binz et al. pose the issue that the chosen prior might not reflect all the functions (and the associated probability). However, similar concerns arise in the context of meta-learned models. What justifications exist for the selection of training data? How does one determine which functions to employ in the tasks used for training the model? Even more, on which tasks the model should be trained? Are these decisions easier to justify from a rational perspective as compared to the Bayesian counterpart? If the definition of the priors is considered a main obstacle of Bayesian inference to large world problems, a similar challenge extends to the decisions mentioned above, which determine the initial parameterization of meta-learned models and could be conceived as equivalent to a "prior" (Griffiths et al., 2019). Finally, if it is the impossibility to "*have access to a prior or a likelihood*" the main obstacle to large world problems, what is "*the unique feature of meta-learned models*" compared to other Bayesian methods that can construct their own empirical priors (e.g., hierarchical models and empirical Bayes; Friston & Stephan, 2007) or that bypass the evaluation of the likelihood function (e.g., approximate Bayesian computation: Beaumont, 2010; likelihood-free inference: Papamakarios, Nalisnick, Rezende, Mohamed, & Lakshminarayanan, 2021; simulation-based inference: Cranmer, Brehmer, & Louppe, 2020)?

Second, it should be noted that the "meta-learning" feature is not exclusive of meta-learned models in machine learning, but it can be achieved using hierarchical Bayesian models (Grant, Finn, Levine, Darrell, & Griffiths, 2018; Griffiths et al., 2019; Kemp, Perfors, & Tenenbaum, 2007; Li, Callaway, Thompson, Adams, & Griffiths, 2023). Hence, if meta-learning is taken as an argument to enable computational models to face large world problems, it cannot be used as an argument in favor of meta-learned models over hierarchical Bayesian inference.

Putting together our first two concerns, we think that a more fair comparison between meta-learned models (as defined in the target article) and hierarchical (approximate) Bayesian models would have been necessary to assert that meta-learned models contain "*unique*" features to address large world problems.

A further concern regards meta-learning as a solution for large world problems. Following Binmore (2007), the distinction between small and large worlds can be interpreted as making decisions under risk or uncertainty, respectively. In the first case, decision makers know all contingencies of the problem and fully apply the Bayes' rule to make the optimal decision. Large world problems are situations characterized by uncertainty about the causes and the likelihood of the events. In other terms, large world problems can be conceived as situations in which environmental assumptions previously acquired do not hold. However, if meta-learned models need to be retrained when environmental assumptions differ from the training, it follows that the use of meta-learned models can be justified only in small worlds, where previous knowledge can be used to make choices.

Finally, it should be highlighted that the target article grounds meta-learned models on the rational analysis framework (Anderson, 1991) given their property of approximate Bayes optimal solutions. However, Savage's and Binmore's argument was that there is no special justification for rational Bayesian solutions to large world problems. In our opinion, if one wants to hold with this rational perspective, neither Bayesian nor meta-learned models can be considered idoneous to model decision making under uncertainty. However, a possible way out of this impasse can come from psychological and cognitive research fields that have investigated decision making under uncertainty. Theoretical frameworks like the free-energy principle (Friston et al., 2023) or reinforcement learning (Dimitrakakis & Ortner, 2022; Kochenderfer, 2015) have investigated how learning under uncertainty occurs and it is used to construct beliefs that guide decisions in situations where causes of the event are unknown. In our opinion, implementing ideas from these frameworks in the models can be a promising way to solve large world problems.

In conclusion, we do not think that Binz et al. have provided convincing support for the claim "*The ability to construct Bayes-optimal learning algorithms for large world problems is a unique feature of the meta-learning framework.*" We suggest that grounding on the rational analysis of cognition framework is not sufficient for modeling decisions in large worlds, and that exploring and integrating other theoretical frameworks could offer valuable insights to advance their research program.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interest. None.

References

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*(3), 471–485.

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, *41*, 379–406.

Binmore, K. (2007). Rational decisions in large worlds. *Annales d'Economie et de Statistique*, *86*, 25–41.

Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, *117*(48), 30055–30062.

Dimitrakakis, C., & Ortner, R. (2022). *Decision making under uncertainty and reinforcement learning: Theory and algorithms* (Vol. 223). Springer Nature.

Friston, K. J., Da Costa, L., Sajid, N., Heins, C., Ueltzhöffer, K., Pavliotis, G. A., & Parr, T. (2023). The free energy principle made simpler but not too simple. *Physics Reports*, *1024*, 1–29.

Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, *159*, 417–458.

Grant, E., Finn, C., Levine, S., Darrell, T., & Griffiths, T. (2018). Recasting gradient-based meta-learning as hierarchical Bayes. arXiv preprint arXiv:1801.08930.

Griffiths, T. L., Callaway, F., Chang, M. B., Grant, E., Krueger, P. M., & Lieder, F. (2019). Doing more with less: Meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, *29*, 24–30.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Kochenderfer, M. J. (2015). *Decision making under uncertainty: Theory and application*. MIT press.

Li, M. Y., Callaway, F., Thompson, W. D., Adams, R. P., & Griffiths, T. L. (2023). Learning to learn functions. *Cognitive Science*, *47*(4), e13262.

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, *22*(1), 2617–2680.

Savage, L. J. (1972). *The foundations of statistics*. Courier Corporation.

Meta-learning modeling and the role of affective-homeostatic states in human cognition

Ignacio Cea[a,b,c*]

[a]Faculty of Religious Sciences and Philosophy, Temuco Catholic University, Temuco, Chile; [b]Center for Research, Innovation and Creation, Temuco Catholic University, Temuco, Chile and [c]Universidad Alberto Hurtado, Faculty of Philosophy and Humanities, Philosophy Department, Santiago, Chile
icea@uct.cl
https://igneocj.wixsite.com/ignacio-cea

*Corresponding author.

doi:10.1017/S0140525X24000098, e149

Abstract

The meta-learning framework proposed by Binz et al. would gain significantly from the inclusion of affective and homeostatic elements, currently neglected in their work. These components are crucial as cognition as we know it is profoundly influenced by affective states, which arise as intricate forms of homeostatic regulation in living bodies.

Binz et al. offer a very promising research program based on meta-learning to advance our understanding of cognition. Nonetheless, their proposal could greatly benefit from integrating affective and homeostatic elements, which have been traditionally underrepresented in cognitive science and are totally absent in their article as well. This integration is predicated on the premise that cognitive processes in general are profoundly influenced by affective states (e.g., emotions and moods), which arise as intricate forms of homeostatic regulation in living organisms.

First, there is ample evidence from psychology and affective neuroscience showing the pervasive influence of affective states on cognitive processes (Cea, 2023; Clore & Schiller, 2018). When someone feels affectively moved by new information, she will more likely attend to it and think about it for an extended duration, compared to neutral information (Manns & Bass, 2016). People in positive moods tend to engage in more creative thinking and learn subjects meaningfully (Ormrod, Anderman, & Anderman, 2019), while those feeling sadness or frustration are prone to a shallower learning (Ahmed, Van der Werf, Kuyper, & Minnaert, 2013). This is related to negative emotions being accompanied by cortisol release and the fight-or-flight response that can suppress the prefrontal cortex, thereby hindering higher cognition (Brackett, 2019). Hence, it is reasonable that emotion regulation abilities are the strongest predictors of academic achievement in high-school students (Di Fabio & Palazzeschi, 2009).

Also, our affective states shape what and how we perceive (Barrett, 2017; Cea & Martínez-Pernía, 2023). According to the affective information principle, our feelings signal the value and urgency of any perceived object (Clore & Schiller, 2018). Moreover, by altering people's affects, researchers can sway perceptions, for example, making a beverage seem appealing or distasteful (Berridge & Winkielman, 2003), and people friendly or mean (Li, Moallem, Paller, & Gottfried, 2007). A key brain area involved is the orbitofrontal cortex, which integrates sensory and affective information, ensuring that our perceptions are always imbued with affect (Barrett & Bar, 2009).

Concerning attention, positive emotions generally broaden it, leading to a global focus (Fredrickson & Branigan, 2005), whereas negative emotions narrow it, fostering detail-oriented processing (Bar-Haim, Lamy, Pergamin, Bakermans-Kranenburg, & Van Ijzendoorn, 2007). Also, emotionally charged stimuli are shown to capture attention more effectively than neutral ones, both in terms of speed and focus (Hajcak, Jackson, Ferri, & Weinberg, 2018). This can cause perceptual interference, making subsequent neutral stimuli less noticeable (Wang, Kennedy, & Most, 2012), a process associated with emotional allocation of attentional resources as indicated by the late positive potential in the parietal lobe (Hajcak et al., 2018).

Concerning memory, William James claimed that "an experience may be so exciting emotionally as almost to leave a scar upon the cerebral tissues" (James, 1890, p. 670). People tend to remember information with emotional significance more easily than neutral material (Phelps & Sharot, 2008). Memories associated with strong emotional arousal are recalled more vividly (Schaefer & Philippot, 2005) and with greater detail in certain respects compared to neutral memories (Kensinger & Schacter, 2018). Neurobiologically, the reciprocal interactions between the amygdala and the hippocampus are considered essential for this (McGaugh, 2013).

Concerning the homeostatic roots of affect, Seth and Barrett have independently suggested that affect arises from the brain's inferences about the causes of internal body signals to regulate physiological states (e.g., sugar levels, heart-beat, etc.), ensuring survival based on past experiences (Barrett, 2017; Barrett & Simmons, 2015; Seth, 2021; Seth & Tsakiris, 2018). Hence, according to them, our feelings would then be expressions of our current and future degree of success or failure in staying alive. In this way, moods and emotions would be intimately linked to our bodily nature and homeostatic needs.

This core idea of feelings being rooted in homeostatic regulation in vulnerable systems has been applied to robotics and artificial intelligence. Man and Damasio (2019) propose a novel class of soft robots that incorporate physical vulnerability and self-regulation akin to living organisms. They hypothesize that this would allow them to develop motivations and evaluations reminiscent of feelings in humans, potentially leading to more intelligent interactions with their environments. Similarly, Bronfman, Ginsburg, and Jablonka (2021) propose that feelings may arise in artificial systems constructed with soft materials through the development of homeostatic, self-preservation mechanisms that could allow them to instantiate an open-ended domain-general form of learning, what they call unlimited associative learning. Finally, Yoshida (2017) introduces a reinforcement learning model where agents learn to survive by regulating critical variables like energy levels, using a reward system rooted in homeostatic principles, leading to adaptive behavior. Importantly, all proposals emphasize that the possibility of engineering sentient artificial systems depends on having vulnerable bodies that, akin to ourselves, need to be constantly sensed and regulated to remain integral, and that this would enhance the machines' cognitive capacities.

To conclude, I would like to suggest some potential benefits of incorporating affective and homeostatic elements into the meta-learning research program: (i) *Enhanced adaptability*: By incorporating these elements into the meta-learning algorithms, the resulting computational models may better simulate the adaptability of human cognition, like the ability to adjust learning strategies to changing environmental and internal states; (ii) *richer contextual understanding*: Incorporating affective-homeostatic elements in learning-to-learn processes can result in a deeper understanding of how emotionally salient contexts influence cognition; (iii) *improved learning efficiency*: Affective-homeostatic signals can guide attention and memory processes, leading to more efficient learning. Meta-learning algorithms that incorporate affective-homeostatic signals or mechanisms may achieve higher efficiency in adapting learning algorithms to new tasks or to new information; (iv) *more realistic simulations*: By incorporating affect and homeostasis, meta-learned models may more accurately simulate human cognition, which is inherently influenced by affective-bodily states; and (v) *cross-domain generalization*: The integration of affective-homeostatic states may facilitate better generalization across different cognitive domains, as affect and bodily regulation often play a role in a wide range of cognitive tasks, from decision making to social interactions.

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

In sum, I encourage Binz et al. to consider the beneficial prospects of incorporating these elements into their proposed research program, so as to acknowledge the intertwined nature of affect, bodily homeostasis, and human cognition.

Acknowledgements. I am very grateful to Kingson Man for his constructive comments on an earlier version of this commentary. I am also very grateful to Thomas Wachter for his encouraging appreciation of my previous work on affect, homeostasis, and cognition.

Financial support. This research was funded by ANID-Fondecyt Postdoctoral grant #3210707 and the Center for Research, Innovation and Creation, Temuco Catholic University.

Competing interest. None.

References

Ahmed, W., Van der Werf, G., Kuyper, H., & Minnaert, A. (2013). Emotions, self-regulated learning, and achievement in mathematics: A growth curve analysis. *Journal of Educational Psychology*, *105*(1), 150.

Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J., & Van Ijzendoorn, M. H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. *Psychological Bulletin*, *133*(1), 1.

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.

Barrett, L. F., & Bar, M. (2009). See it with feeling: Affective predictions during object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364* (1521), 1325–1334.

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, *16*(7), 419–429.

Berridge, K., & Winkielman, P. (2003). What is an unconscious emotion?(The case for unconscious "liking"). *Cognition and Emotion*, *17*(2), 181–211. https://doi.org/10.1080/02699930302289

Brackett, M. (2019). *Permission to feel: Unlocking the power of emotions to help our kids, ourselves, and our society thrive*. Celadon Books.

Bronfman, Z., Ginsburg, S., & Jablonka, E. (2021). When will robots be sentient? *Journal of Artificial Intelligence and Consciousness*, *8*(02), 183–203.

Cea, I. (2023). The somatic roots of affect: Toward a body-centered education. In P. Fossa, & C. Cortés-Rivera (Eds.), *Affectivity and learning: Bridging the gap between neurosciences, cultural and cognitive psychology* (pp. 555–583). Springer.

Cea, I., & Martínez-Pernía, D. (2023). Continuous organismic sentience as the integration of core affect and vitality. *Journal of Consciousness Studies*, *30*(3–4), 7–33. https://doi.org/https://doi.org/10.53765/20512201.30.3.007

Clore, G., & Schiller, A. (2018). New light on the affect-cognition connection. In L. F. Barrett, M. Lewis, & J. M. Haviland-Jones (Eds.), *Handbook of emotions*, (4th ed., pp. 532–546). The Guilford Press.

Di Fabio, A., & Palazzeschi, L. (2009). An in-depth look at scholastic success: Fluid intelligence, personality traits or emotional intelligence? *Personality and Individual Differences*, *46*(5–6), 581–585.

Fredrickson, B. L., & Branigan, C. (2005). Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition & Emotion*, *19*(3), 313–332.

Hajcak, G., Jackson, F., Ferri, J., & Weinberg, A. (2018). Emotion and attention. In L. F. Barrett, M. Lewis, & J. M. Haviland-Jones (Eds.), *Handbook of emotions*, (4th ed., pp. 595–609). The Guilford Press.

James, W. (1890). *The principles of psychology* (Vol. 1, Issue n/a). Dover Publications.

Kensinger, E. A., & Schacter, D. L. (2018). Memory and emotion. In L. F. Barrett, M. Lewis, & J. M. Haviland-Jones (Eds.), *Handbook of emotions*, (4th ed., pp. 564–578). The Guilford Press.

Li, W., Moallem, I., Paller, K. A., & Gottfried, J. A. (2007). Subliminal smells can guide social preferences. *Psychological Science*, *18*(12), 1044–1049.

Man, K., & Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, *1*(10), 446–452.

Manns, J. R., & Bass, D. I. (2016). The amygdala and prioritization of declarative memories. *Current Directions in Psychological Science*, *25*(4), 261–265.

McGaugh, J. L. (2013). Making lasting memories: Remembering the significant. *Proceedings of the National Academy of Sciences*, *110*(supplement_2), 10402–10407.

Ormrod, J. E., Anderman, E. M., & Anderman, L. H. (2019). *Educational psychology: Developing learners*, (10th ed.). Pearson.

Phelps, E. A., & Sharot, T. (2008). How (and why) emotion enhances the subjective sense of recollection. *Current Directions in Psychological Science*, *17*(2), 147–152.

Schaefer, A., & Philippot, P. (2005). Selective effects of emotion on the phenomenal characteristics of autobiographical memories. *Memory*, *13*(2), 148–160.

Seth, A. (2021). *Being you: A new science of consciousness*. Penguin.

Seth, A. K., & Tsakiris, M. (2018). Being a beast machine: The somatic basis of selfhood. *Trends in Cognitive Sciences*, *22*(11), 969–981. https://doi.org/10.1016/j.tics.2018.08.008

Wang, L., Kennedy, B. L., & Most, S. B. (2012). When emotion blinds: A spatiotemporal competition account of emotion-induced blindness. *Frontiers in Psychology*, *3*, 438.

Yoshida, N. (2017). Homeostatic agent for general environment. *Journal of Artificial General Intelligence*, *8*(1), 1.

Quantum Markov blankets for meta-learned classical inferential paradoxes with suboptimal free energy

Kevin B. Clark[a–k,*]

[a]Cures Within Reach, Chicago, IL, USA; [b]Felidae Conservation Fund, Mill Valley, CA, USA; [c]Campus and Domain Champions Program, Multi-Tier Assistance, Training, and Computational Help (MATCH) Track, National Science Foundation's Advanced Cyberinfrastructure Coordination Ecosystem: Services and Support (ACCESS); [d]Expert Network, Penn Center for Innovation, University of Pennsylvania, Philadelphia, PA, USA; [e]Network for Life Detection (NfoLD), NASA Astrobiology Program, NASA Ames Research Center, Mountain View, CA, USA; [f]Multi-Omics and Systems Biology & Artificial Intelligence and Machine Learning Analysis Working Groups, NASA GeneLab, NASA Ames Research Center, Mountain View, CA, USA; [g]Frontier Development Lab, NASA Ames Research Center, Mountain View, CA, USA; [h]SETI Institute, Mountain View, CA, USA; [i]Peace Innovation Institute, The Hague 2511, Netherlands & Stanford University, Palo Alto, CA, USA; [j]Shared Interest Group for Natural and Artificial Intelligence (sigNAI), Max Planck Alumni Association, Berlin, Germany and [k]Biometrics and Nanotechnology Councils, Institute for Electrical and Electronics Engineers (IEEE), New York, NY, USA
kbclarkphd@yahoo.com
www.linkedin.com/pub/kevin-clark/58/67/19a
https://access-ci.org/

*Corresponding author.

doi:10.1017/S0140525X24000244, e150

Abstract

Quantum active Bayesian inference and quantum Markov blankets enable robust modeling and simulation of difficult-to-render natural agent-based classical inferential paradoxes interfaced with task-specific environments. Within a non-realist cognitive completeness regime, quantum Markov blankets ensure meta-learned irrational decision making is fitted to explainable manifolds at optimal free energy, where acceptable incompatible observations or temporal Bell-inequality violations represent important verifiable real-world outcomes.

Applying a rational analysis framework of cognition, Binz et al. resolutely embrace the escalating use of meta-learning to re-construct Bayes optimal-learning algorithms and solutions to explain the metaphysical relationship of mind to environment. Such logicomathematical descriptions of cognition non-trivially approximate properties of mind, including agent-based decision processes, to that of ideal statistical inference and the structure of natural tasks and environments. The authors nonetheless fail to satisfactorily introduce Markov blankets, which would expand their cognitive construct beyond standard applications of analytical and numerical tools to demark relations represented in directed Bayesian networks or graphs and variational probabilistic inference. Markov or Pearl blankets, coined by Bruineberg, Dolega, Dewhurst, and Baltieri (2021) to acknowledge the original decades-old epistemic concept developed by Pearl (1988), earned deserving attention for their utility in decision sciences, offering tractable optimization methods to identify, partition, and understand groups of marginally and conditionally (in)dependent variables associated with complex systems and causal predictions and attributions. In more recent years, however, a trend started by Friston and coworkers (e.g., Hipólito et al., 2021; Friston et al., 2021; Ramstead, Badcock, & Friston, 2018) promotes extension of Markov blankets within a free-energy, active-inference framework to describe the physical interface and reciprocal interactions between agents and their environments. These so-called Friston blankets, a term also coined by Bruineberg et al. (2021) to differentiate their usage from Pearl blankets, cleverly articulate embedded philosophical axioms of mind, body, and environment without fully capturing the logicomathematical rigor and cogency found in typical uses of Pearl blankets and Bayesian inference. Taking a reasonably conventional position on the state-of-art of Friston blankets, theorists supportive of free-energy models must either accept Markov blankets as formal technical innovations that enable practical worthwhile science in absence of known compelling philosophical conclusions about the nature of cognition and life or as dubious mathematics-driven metaphysics interpretations of reality with promising high-impact implications for elucidating cognition and life upon separate experimental biophysical validation. The merits of such a statement seemly convey a fair, albeit critical, edict to the scientific community – one that perhaps discouraged Binz et al. from clarifying Markov blankets and leaves the impression that good productive science enlisting Markov blankets as instruments for inference returns only theoretically mundane findings about cognition and, possibly, life. But, that is not the case and the authors' well-structured arguments may fool readers into believing that this position is true since Binz et al. disappointingly content themselves with only discussing classical Bayesian inference and non-paradoxical cognition, maybe because Friston and coworkers also never stray beyond a classical formulation of their free-energy principle for active inference.

Weaknesses in Binz et al.'s narrow perspective on meta-leaning and cognition may be contrasted and reshaped by exciting findings produced with quantum decision theory, a strict valid quantum-statistical approach capable of defining probabilistic human inference unconfined by the physical mechanical world (Aerts & Aerts, 1995; Aerts, Broakaert, Gabara, & Sozzo, 2016; Ashtiani & Azgomi, 2015; Busemeyer & Bruza, 2011; Busemeyer, Wang, & Lambert-Mogiliansky, 2009; Clark, 2011, 2012, 2014b, 2015, 2017; Pinto Moreira, Fell, Dehdashti, Bruza, & Wichert, 2020; Pothos & Busemeyer, 2013). Quantum decision theory, with the aid of quantum networks or graphs, quantum Bayesian inference, and other computational features, demonstrates robust successes in modeling and simulating difficult-to-render cognitive phenomena overlooked by Binz et al., especially causal judgment errors or paradoxes unaccounted for by classical decision theory, including conjunctive and disjunctive fallacies, the Allais paradox, and the Ellsberg or planning paradox (Atmanspacher & Römer, 2012; Busemeyer, Pothos, Franco, & Trueblood, 2011; Clark, 2021b; Favre, Wittwer, Heinimann, Yukalov, & Sornette, 2016; Moreira & Wichert, 2016a, 2016b, 2018b; Pothos & Busemeyer, 2009, 2013). The great explanatory power of quantum decision networks (Bianconi, 2002a, 2002b, 2003; Bianconi & Barabási, 2001; Li, Iqbal, Perc, Chen, & Abbott, 2013) permits expression of quantum Markov chains and blankets (Brandao, Piani, & Horodecki, 2015; Moreira & Wichert, 2018a; Qi & Ranard, 2021; Sutter, 2018; Wichert, Pinto Moreira, & Bruza, 2020), which bound suboptimal free-energy classical inferential paradoxes from optimal free-energy quantum inferential solutions. In this context of meta-learned cognitive inference, both quantum Bayesian inference and Markov blankets provide epistemic tools, analogous to variational Bayesian inference and Pearl blankets, to legitimately approximate irrational human decision making and cognition within a cognitive-completeness constraint. That constraint, not considered by Binz et al., strongly limits degrees of freedom for emergence of meta-learned subjective physical reality (cf. Blume-Kohout & Zurek, 2006; Brandao et al., 2015; Clark, 2014a, 2017, 2019, 2020, 2021b, 2023; Yearsley & Pothos, 2014), a scenario which helps affirm the idea that Markov blankets may yield meaningful philosophical conclusions regarding the nature of cognition and life.

Cognitive completeness (Tressoldi, Maier, Buechner, & Khrennikov, 2015; Yearsley & Pothos, 2014) encapsulates a black-box approach that isolates any studied cognitive system from the formidable environment-significant measurement problem of quantum mechanics. Some scientists insist the scalable neurophysiological contents of this black box map onto classical or quantum cognitive states relevant to particular sets of respective rational or irrational decisions and their corresponding outcome probabilities (Clark, 2017, 2021a; Wang & Busemeyer, 2015; Yearsley & Pothos, 2014). For example, if one abandons the constraint of cognitive realism – the assertion that all (meta-learned) cognitive events emerge from classical deterministic neurophysiology – then cognitive systems and their decisional outcomes may be completely described by dimensions or sets of similarity classes on the set of all probability distributions over deterministic and indeterminate (or stochastic) neuropsychological variables and associated environmental settings. Further, arguably more fundamental cognitive completeness and Markov blanket partitions imply non-disturbing measurements should be non-invasive on brain physiology and cognition, with disturbing measurements affecting outcomes of hidden neurophysiological and cognitive variables in a quantum-sensitive manner. Disturbing measurements violate temporal Bell or Leggett–Garg inequalities, signifying violations of classical (meta-learned) cognition, such as Markov-blanketed inferential paradoxes at suboptimal free energy. Restructuring these cognitive phenomena within quantum decision theory and quantum Markov blankets creates opportunities for irrational decision making to be organized into explainable manifolds at optimal free energy, tantamount to quantum active Bayesian inference where observable incompatibility or inequality violations are acceptable (Clark, 2021a). Such theoretical elegance in describing complex cognition forces classical and quantum aspects of brain structure and function into a newer realm of logicomathematical formalism with verifiable, important real-world consequences.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interest. None.

References

Aerts, D., & Aerts, S. (1995). Applications of quantum statistics in psychological studies of decision processes. *Foundations of Science*, *1*, 85–97.

Aerts, D., Broakaert, J., Gabara, L., & Sozzo, S. (Eds.) (2016). *Quantum structures in cognitive and social sciences*. Frontiers Research Topics Ebook. Frontiers Media SA. ISBN 978-2-88919-876-4.

Ashtiani, M., & Azgomi, M. A. (2015). A survey of quantum-like approaches to decision making and cognition. *Mathematical Social Sciences*, *75*, 49–80.

Atmanspacher, H., & Römer, H. (2012). Order effects in sequential measurements of noncommuting psychological observables. *Journal of Mathematical Psychology*, *56*, 274–280.

Bianconi, G. (2002a). Growing Cayley trees described by a Fermi distribution. *Physical Review E: Statistical, Nonlinear, Soft Matter Physics*, *66*, 036116.

Bianconi, G. (2002b). Quantum statistics in complex networks. *Physical Review E: Statistical, Nonlinear, Soft Matter Physics*, *66*, 056123.

Bianconi, G. (2003). Size of quantum networks. *Physics Review E*, *67*(5, Pt 2), 056119.

Bianconi, G., & Barabási, A.-L. (2001). Bose-Einstein condensation in complex networks. *Physics Review Letters*, *86*, 5632–5635.

Blume-Kohout, R., & Zurek, W. H. (2006). Quantum Darwinism: Entanglement, branches, and the emergent classicality of redundantly stored quantum information. *Physical Review A*, *73*(6), 062310.

Brandao, F. G. S. L., Piani, M., & Horodecki, P. (2015). Generic emergence of classical features in quantum Darwinism. *Nature Communications*, *6*, 7908.

Bruineberg, J., Dolega, K., Dewhurst, J., & Baltieri, M. (2021). The emperor's new Markov blankets. *Behavior and Brain Sciences*, *45*, e183.

Busemeyer, J. R., & Bruza, P. (2011). *Quantum models of cognition and decision making*. Cambridge University Press. ISBN 13 978-1107419889.

Busemeyer, J. R., Pothos, E., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment "errors". *Psychological Reviews*, *118*, 193–218.

Busemeyer, J. R., Wang, Z., & Lambert-Mogiliansky, A. (2009). Empirical comparison of Markov and quantum models of decision making. *Journal of Mathematical Psychology*, *53*, 423–433.

Clark, K. B. (2011). Live soft-matter quantum computing. In E. C. Salander (Ed.), *Computer search algorithms* (pp. 1–24). Nova Science Publishers, Inc. ISBN 978-1-61122-527-3.

Clark, K. B. (2012). A statistical mechanics definition of insight. In A. G. Floares (Ed.), *Computational intelligence* (pp. 139–162). Nova Science Publishers, Inc. ISBN 978-1-62081-901-2.

Clark, K. B. (2014a). Basis for a neuronal version of Grover's quantum algorithm. *Frontiers in Molecular Neuroscience*, *7*, 29.

Clark, K. B. (2014b). Evolution of affective and linguistic disambiguation under social eavesdropping pressures. *Behavioral and Brain Sciences*, *37*(6), 551–552.

Clark, K. B. (2015). Insight and analysis problem solving in microbes to machines. *Progress in Biophysics and Molecular Biology*, *119*, 183–193.

Clark, K. B. (2017). Cognitive completeness of quantum teleportation and superdense coding in neuronal response regulation and plasticity. *Proceedings of the Royal Society B: Biological Sciences*. eLetter. https://royalsocietypublishing.org/doi/suppl/10.1098/rspb.2013.3056

Clark, K. B. (2019). Unpredictable homeodynamic and ambient constraints on irrational decision making of aneural and neural foragers. *Behavioral and Brain Sciences*, *42*, e40.

Clark, K. B. (2020). Digital life, a theory of minds, and mapping human and machine cultural universals. *Behavioral and Brain Sciences*, *43*, e98.

Clark, K. B. (2021a). Quantum decision corrections for the neuroeconomics of irrational movement control and goal attainment. *Behavioral and Brain Sciences*, *44*, e127.

Clark, K. B. (2021b). Classical and quantum information processing in aneural to neural cellular decision making on Earth and perhaps beyond. *Bulletin of the American Astronomical Society*, *53*(4), 33.

Clark, K. B. (2023). Neural field continuum limits and the partitioning of cognitive-emotional brain networks. *Biology*, *12*(3), 352.

Favre, M., Wittwer, A., Heinimann, H. R., Yukalov, V. I., & Sornette, D. (2016). Quantum decision theory in simple risky choices. *PLoS ONE*, *11*(12), e0168045.

Friston, K. J., Parr, T., Hipolito, I., Magrou, L., & Razi, A. (2021). Parcels and particles: Markov blankets in the brain. *Network Neuroscience*, *5*(1), 211–251.

Hipolito, I., Ramstead, M. J. D., Convertino, L., Bhat, A., Friston, K., & Parr, T. (2021). Markov blankets in the brain. *Neuroscience and Biobehavioral Reviews*, *125*, 88–97.

Li, Q., Iqbal, A., Perc, M., Chen, M., & Abbott, D. (2013). Coevolution of quantum and classical strategies on evolving random networks. *PLoS ONE*, *8*(7), e68423.

Moreira, C., & Wichert, A. (2016a). Quantum-like Bayesian networks for modeling decision making. *Frontiers in Psychology*, *7*, 11.

Moreira, C., & Wichert, A. (2016b). Quantum probabilistic models revisited: The case of disjunction effects in cognition. *Frontiers in Psychology*, *4*, 26.

Moreira, C., & Wichert, A. (2018a). Are quantum-like Bayesian networks more powerful than classical Bayesian networks? *Journal of Mathematical Psychology*, *28*, 73–83.

Moreira, C., & Wichert, A. (2018b). Introducing quantum-like influence diagrams for violations of the Sure Thing Principle. *International symposium on quantum interactions* Q1 2018: 91–108.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann Publishers Inc. ISBN 1558604790.

Pinto Moreira, C., Fell, L., Dehdashti, S., Bruza, P., & Wichert, A. (2020). Towards a quantum-like cognitive architecture for decision-making. *Behavioral and Brain Sciences*, *43*, e171–e178.

Pothos, E. M., & Busemeyer, J. R. (2009). A quantum probability explanation for violations of "rational" decision theory. *Proceedings of Biological Sciences*, *276*, 2171–2178.

Pothos, E. M., & Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences*, *36*, 255–327.

Qi, X.-L., & Ranard, D. (2021). Emergent classicality in general multipartite states and channels. *Quantum*. https://arxiv.org/abs/2001.01507

Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2018). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews*, *24*, 1–16.

Sutter, D. (2018). *Approximate quantum Markov chains*. Springer. ISBN 978-3-319-78732-9.

Tressoldi, P. E., Maier, M. A., Buechner, V. L., & Khrennikov, A. (2015). A macroscopic violation of no-signaling in time inequalities? How to test temporal entanglement with behavioral observables. *Frontiers in Psychology*, *6*, 1061.

Wang, Z., & Busemeyer, J. (2015). Reintroducing the concept of complementarity to psychology. *Frontiers in Psychology*, *6*, 1822.

Wichert, A., Pinto Moreira, C., & Bruza, P. (2020). Balanced quantum-like Bayesian networks. *Entropy*, *22*(2), 1701–1726.

Yearsley, J. M., & Pothos, E. M. (2014). Challenging the classical notion of time in cognition: A quantum perspective. *Proceedings of the Royal Society London B: Biological Sciences*, *281*, 20133056.

Meta-learning goes hand-in-hand with metacognition

Chris Fields[a] and James F. Glazebrook[b,c]*

[a]Allen Discovery Center, Tufts University, Medford, MA, USA; [b]Department of Mathematics and Computer Science, Eastern Illinois University, Charleston, IL, USA and [c]Adjunct Faculty (Mathematics), University of Illinois at Urbana-Champaign, Urbana, IL, USA
fieldsres@gmail.com
jfglazebrook@eiu.edu
https://chrisfieldsresearch.com
https://faculy.math.illinois.edu/glazebro/

*Corresponding author.

doi:10.1017/S0140525X24000256, e151

Abstract

Binz et al. propose a general framework for meta-learning and contrast it with built-by-hand Bayesian models. We comment on some architectural assumptions of the approach, its relation to the active inference framework, its potential applicability to living systems in general, and the advantages of the latter in addressing the explanation problem.

Binz et al. craft a comprehensive outline for advancing meta-learning (MetaL) on the basis of several arguments concerning the tractability of optimal learning algorithms, manipulation of complexity, and integration into the rational aspects of cognition, all seen as basic requirements for a domain-general model of cognition. Architectural features include an inductive process from experience driven by repetitive interaction with the environment, necessitating (i) an inner loop of "base learning," and (ii) an outer loop (or MetaL) process through which the system is effectively trained by the environment to ameliorate its inner loop learning algorithms. A key aspect of the model is its dependence on the relation between the typical duration of a (general, MetaL) problem-solving episode and the typical duration of a (particular, learned) solution.

While Binz et al. focus on MetaL as a practical methodology for modeling human cognition, it is also interesting to ask how MetaL as Binz et al. describe it, fits into the conceptual framework of cognition in general, and also to ask how it applies both to organisms other than humans and to artificial (or hybrid) systems operating in task environments very different from the human task environment. From a broad perspective, MetaL is one function of metacognition (e.g., Cox, 2005; Flavell, 1979; Shea & Frith, 2019). Both MetaL and metacognition more generally engage memory and attention as they are neurophysiologically enacted by brain regions including the default mode network (Glahn et al., 2010), as reviewed for the two theories in Wang (2021) and Kuchling, Fields, and Levin (2022), respectively.

When MetaL is viewed as implemented by a metaprocessor that is a proper component of a larger cognitive system, one can ask explicitly about the metaprocessor's task environment and how it relates to the larger system's task environment. MetaL operates in a task environment of learning algorithms and outcomes, or equivalently, a task environment of metaparameters and test scores. How the latter are measured is straightforward for a human modeler employing MetaL as a methodology, but is less straightforward when an explicit system-scale architecture must be specified. The question in this case becomes that of how the object-level components of a system use the feedback received from the external environment to train the metaprocessor. The answer cannot, on pain of infinite regress, be MetaL. The relative inflexibility of object-level components as "trainers" of their associated metaprocessors effectively bakes in some level of non-optimality in any multilayer system.

Binz et al. emphasize that MetaL operates on a longer timescale than object-level learning. Given a task environment that imposes selective pressures with different timescales, natural selection will drive systems toward layered architectures that exhibit MetaL (Kuchling et al., 2022). Indeed the need for a "learning to learn" capability has long been emphasized in the active-inference literature (e.g., Friston et al., 2016). Active inference under the free-energy principle (FEP) is in an important sense "just physics" (Friston, 2019; Friston et al., 2023; Ramstead et al., 2022); indeed the FEP itself is just a classical limit of the principle of unitarity, that is, of conservation of information (Fields et al., 2023; Fields, Friston, Glazebrook, & Levin, 2022). One might expect, therefore, that MetaL as defined by Binz et al. is not just useful, but ubiquitous in physical systems with sufficient degrees of freedom. As this is at bottom a question of mathematics, testing it does not require experimental investigation.

What does call out for experimental investigation is the extent to which MetaL can be identified in systems much simpler than humans. Biochemical pathways can be trained, via reinforcement learning, to occupy different regions of their attractor landscapes (Biswas, Manika, Hoel, & Levin, 2021, 2022). Do sufficiently complex biochemical networks that operate on multiple timescales exhibit MetaL? Environmental exploration and learning are ubiquitous throughout phylogeny (Levin, 2022, 2023); is MetaL equally ubiquitous? Learning often amounts to changing the salience distribution over inputs, or in Bayesian terms, adjusting precision assignments to priors. To what extent can we describe the implementation of MetaL by organisms in terms of adjustments of sensitivity/salience landscapes – and hence attractor landscapes – on the various spaces that compose their *umwelts*?

As Binz et al. point out, in the absence of a mechanism for concrete mathematical analysis, MetaL forsakes interpretable analytic solutions and hence generates an "explanation problem" (cf. Samek, Montavon, Lapuschkin, Anders, & Müller, 2021). As in the case of deep AI systems more generally, experimental techniques from cognitive psychology may be the most productive approach to this problem for human-like systems (Taylor & Taylor, 2021). Relevant to this is an associated spectrum of ideas, including how problem solving is innately perceptual, how inference is "Bayesian satisficing" not optimization (Chater, 2018; Sanborn & Chater, 2016), the relevance of heuristics (Gigerenzer & Gaissmaier, 2011; cf. Fields & Glazebrook, 2020), and how heuristics, biases, and confabulation limit reportable self-knowledge (Fields, Glazebrook, & Levin, 2024). Here again, the possibility of studying MetaL in more tractable experimental systems in which the implementing architecture can be manipulated biochemically and bioelectrically, may offer a way forward not available with either human subjects or deep neural networks.

Financial support. The authors have received no funding towards this contribution.

Competing interests. None.

References

Biswas, S., Clawson, W., & Levin, M. (2022). Learning in transcriptional network models: Computational discovery of pathway-level memory and effective interventions. *International Journal of Molecular Sciences*, *24*, 285.

Biswas, S., Manika, S., Hoel, E., & Levin, M. (2021). Gene regulatory networks exhibit several kinds of memory: Quantification of memory in biological and random transcriptional networks. *IScience*, *24*, 102131.

Chater, N. (2018). *The mind is flat. The remarkable shallowness of the improvising brain*. Yale University Press.

Cox, M. T. (2005). Metacognition in computation: A selected research review. *Artificial Intelligence*, *169*, 104–141.

Fields, C., Fabrocini, F., Friston, K. J., Glazebrook, J. F., Hazan, H., Levin, M., & Marcianò, A. (2023). Control flow in active inference systems, part I: Classical and quantum formulations of active inference. *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, *9*, 235–245.

Fields, C., Friston, K. J., Glazebrook, J. F., & Levin, M. (2022). A free energy principle for generic quantum systems. *Progress in Biophysics and Molecular Biology*, *173*, 36–59.

Fields, C., & Glazebrook, J. F. (2020). Do process-1 simulations generate the epistemic feelings that drive process-2 decision making? *Cognitive Processing*, *21*, 533–553.

Fields, C., Glazebrook, J. F., & Levin, M. (2024). Principled limitations on self-representations for generic physical systems. *Entropy*, *26*(3), 194.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*, 906.

Friston, K. J. (2019). A free energy principle for a particular physics, Preprint arxiv:1906.10184.

Friston, K. J., Da Costa, L., Sakthivadivel, D. A. R., Heins, C., Pavliotis, G. A., Ramstead, M. J., & Parr, T. (2023) Path integrals, particular kinds, and strange things. *Physics of Life Reviews*, *47*, 35–62.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews*, *68*, 862–879.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Glahn, D. C., Winkler, A. M., Kochunov, P., & Blangero, J. (2010) Genetic control over the resting brain. *Proceedings of the National Academy of Sciences of the USA*, *10*(7), 1223–1228.

Kuchling, F., Fields, C., & Levin, M. (2022). Metacognition as a consequence of competing evolutionary time scales. *Entropy*, *24*, 601.

Levin, M. (2022). Technological approach to mind everywhere: An experimentally-grounded framework for understanding diverse bodies and minds. *Frontiers in Systems Neuroscience*, *16*, 768201.

Levin, M. (2023). Darwin's agential materials: Evolutionary implications of multiscale competency in developmental biology. *Cellular and Molecular Life Sciences*, *80*(6), 142.

Ramstead, M. J.,Sakthivadivel, D. A. R., Heins, C., Koudahl, M., Millidge, B., Da Costa, L., … Friston, K. J. (2022). On Bayesian mechanics: A physics of and by beliefs. *Interface Focus*, *13*(2923), 20220029.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021) Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, *109*, 247–278.

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893.
Shea, N., & Frith, C. D. (2019). The global workspace needs metacognition. *Trends in Cognitive Sciences*, *23*, 560–571.
Taylor, J. E. T., & Taylor, G. W. (2021). Artificial cognition: How experimental psychology can help generate artificial intelligence. *Psychonomic Bulletin and Review*, *28*, 454–475.
Wang, J. X. (2021). Meta-learning in artificial and natural intelligence. *Current Opinion in Behavioral Sciences*, *38*, 90–95.

The meta-learning toolkit needs stronger constraints

Erin Grant*

UCL Gatsby Unit, Sainsbury Wellcome Centre, University College London, London, UK
erin.grant@ucl.ac.uk
https://eringrant.github.io/

*Corresponding author.

doi:10.1017/S0140525X24000104, e152

Abstract

The implementation of meta-learning targeted by Binz et al. inherits benefits and drawbacks from its nature as a connectionist model. Drawing from historical debates around bottom-up and top-down approaches to modeling in cognitive science, we should continue to bridge levels of analysis by constraining meta-learning and meta-learned models with complementary evidence from across the cognitive and computational sciences.

Meta-learning as a model allows researchers to posit how human and other biological learning systems might learn from experience in a structured manner, including by relating experiences across timescales or latent causes non-uniformly. Meta-learning as a tool allows researchers to posit flexible and data-driven learning algorithms as computational models of human learning than those are readily expressed by machine learning algorithms such as gradient descent with canonical parameters, or inference in a Bayesian model in which exact inference is tractable. These senses of a "meta-learning" and a "meta-learned" model align with the dichotomy employed in Binz et al.

Meta-learning in both senses and using the implementation focused on in Binz et al. – a recurrent neural network – further inherits characteristics of connectionism: Universal approximation, ease of specification, manipulability (including of complexity), and integration of neuroscientific findings, which Binz et al. rightly note as positives. However, this implementation of meta-learning also inherits the challenges of a connectionist approach: Lack of interpretability (the ease with which humans can understand the workings and outputs of a system) and controllability (the ability to modify a model's behavior or learning process to achieve specific outcomes).

These benefits and drawbacks of the bottom-up, emergentist approach of connectionism have been discussed at length, including in this journal (Smolensky, 1988). As a result of these discussions, a common ground between these and top-down structured approaches such as Bayesian cognitive modeling has emerged: That models posed in different description languages may *not* be at odds simply because they are posed at different levels of analysis, and in fact *should* be tested for complementarity (Rogers & McClelland, 2008; Griffiths, Vul, & Sanborn, 2012).

It is this integrative approach that I view as the most fruitful in examining the validity of meta-learning and meta-learned cognitive models precisely because (1) it allows us to address the challenges of working within a single paradigm (say, the lack of interpretability of a connectionist approach) at the same time as (2) providing stronger grounds on which to refute a cognitive model (say, by its inconsistency with evidence from neural recordings, or its inability to account for how an ecological task is solved). Making use of the former benefit is especially critical, as the meta-learned models commonly employed, including by Binz et al., have the potential to be even more inscrutable than a connectionist model initialized in a data-agnostic way.

Binz et al. discuss two studies of meta-learning and meta-learned models that bridge levels of analysis in this manner: Firstly, a meta-learning algorithm has been tested against experimental neuroscience findings in prefrontal cortex (Wang et al., 2018). Secondly, a meta-learned recurrent neural network can approximate the posterior predictive distribution picked out as optimal by a Bayesian approach (Ortega et al., 2019). Connecting neuroscientific findings with computational-level analysis via algorithm is an exciting result. However, as Binz et al. note, the goodness of fit of the meta-learned approximation employed in both studies is not guaranteed, and has been empirically demonstrated to be poor.

As a contrast to an approach that makes use of approximation, our work (Grant, Finn, Levine, Darrell, & Griffiths, 2018) draws a formal connection between a connectionist implementation of meta-learning and inference in a hierarchical Bayesian model by making precise the prior, likelihood, and parameter estimation procedure implied in the use of the meta-learning implementation. Equivalently, this result describes a way to implement a rational solution to a problem of learning-to-learn in a connectionist architecture (though there are likely to be many equivalent implementations). A formal integration across levels like this is tighter than an approximation approach, and therefore provides a firmer footing for integrative constraints across levels of analysis.

Follow-up investigations have made use of this connection between computational-level and algorithmic-level approaches. For example, in McCoy, Grant, Smolensky, Griffiths, and Linzen (2020), we used an analogous setup to Grant et al. (2018) to meta-learn a syllable typology in a limited data setting akin to an impoverished language learning environment. To better accommodate the complex dynamics of learning, we relaxed some constraints on the meta-learning algorithm, thus for the moment doing away with the tight connection between the algorithmic and computational levels. However, in sticking with methods – namely tuning the gradient-based initialization for learning in a neural network – for which ongoing research in machine learning is formally characterizing how prior knowledge (Dominé, Braun, Fitzgerald, & Saxe, 2023), including data-driven prior knowledge (Lindsey & Lippl, 2023), interacts with the learning algorithm and environment, my view is that these approaches will soon benefit from tighter connections between the algorithmic and computational levels echoing to the connection derived in Grant et al. (2018).

Absent these connections, because meta-learning and meta-learned models are underconstrained and data-driven, it is challenging to evaluate the validity and implications of these models for our understanding of how experience shapes learning. Thus, scientists interested in the place of meta-learning and meta-learned models in cognitive science should work to make precise the constraints that these models imply across levels of analysis, including by making use of analytical techniques from machine learning, at the same time looking into complementary constraints from experimental neuroscience, and ecologically relevant environments. Given that so many aspects remain open, it is an exciting time to be working with and on meta-learning toolkit.

Acknowledgments. N/A.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

Dominé, C. C., Braun, L., Fitzgerald, J. E., & Saxe, A. M. (2023). Exact learning dynamics of deep linear networks with prior knowledge. *Journal of Statistical Mechanics: Theory and Experiment*, *2023*(11), 114004. https://doi.org/10.1088/1742-5468/ad01b8

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263–268. https://doi.org/10.1177/0963721412447619

Lindsey, J. W., & Lippl, S. (2023). Implicit regularization of multi-task learning and finetuning in overparameterized neural networks. arXiv preprint arXiv:2310.02396. https://doi.org/10.48550/arXiv.2310.02396

McCoy, R. T., Grant, E., Smolensky, P., Griffiths, T. L., & Linzen, T. (2020). Universal linguistic inductive biases via meta-learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. https://doi.org/10.48550/arXiv.2006.16324

Ortega, P. A., Wang, J. X., Rowland, M., Genewein, T., Kurth-Nelson, Z., Pascanu, R., … Legg, S. (2019). Meta-learning of sequential strategies. arXiv preprint arXiv:1905.03030. https://doi.org/10.48550/arXiv.1905.03030

Rogers, T. T., & McClelland, J. L. (2008). A simple model from a powerful framework that spans levels of analysis. *Behavioral and Brain Sciences*, *31*(6), 729–749. doi:10.1017/S0140525X08006067

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*(1), 1–23. https://doi.org/10.1017/S0140525X00052432

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., … Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience 21*, 860–868. https://doi.org/10.1038/s41593-018-0147-8

Learning and memory are inextricable

Sue Llewellyn*

University of Manchester, Manchester, UK
sue.llewellyn@manchester.ac.uk
https://www.humanities.manchester.ac.uk/

*Corresponding author.

doi:10.1017/S0140525X2400013X, e153

Abstract

The authors' aim is to build "more biologically plausible learning algorithms" that work in naturalistic environments. Given that, first, human learning and memory are inextricable, and, second, that much human learning is unconscious, can the authors' first research question of how people improve their learning abilities over time be answered without addressing these two issues? I argue that it cannot.

Learning is the process of acquiring a memory; learning and memory both depend, fundamentally, on association (Fuster, 1999). The inextricability of learning and memory originates because any to-be-learned information is encoded in memory through association, the same associative encoding also enables retrieval from memory (Brown & Craik, 2000; Tulving & Thomson, 1973). Once retrieved, the encoded memory will drive expectations in the same (or similar) environments to those in which the learning took place.

The authors identify their meta-learned models as ones that "acquire their inductive biases from experience, i.e. by repeatedly interacting with an environment." This implies that learning is cumulative over time as the model repeatedly samples the environment. For human learning to be cumulative across time, the learned information must be encoded and retained in memory. Later in the paper, the authors, briefly, acknowledge the contribution of episodic memory and the hippocampus but do not spell out how hippocampal memory systems impact their models.

In relation to the authors' first research question, people can improve their learning abilities through *elaborate* encoding which creates associations between the to-be-learnt episodic material and information already encoded in episodic memory networks (Foer, 2011; Llewellyn, 2013; Yates, 1966). The hippocampus is crucial to both associative encoding and later retrieval of episodes through association (Davachi & Wagner, 2002; Llewellyn, 2013; Thakral, Benoit, & Schacter, 2017). The sequential, associative nature of elements of learned, encoded and retained episodic memories gives rise to rational, step-by-step learning. However, during rapid eye movement (REM) sleep, the hippocampus may take *elements of different* episodic experiences to identify more elaborative, associative, probabilistic and meaningful patterns in experience which may be expressed, visually, as dreams and instantiated as cortical nodes/junctions during non-rapid eye movement (NREM) sleep (Llewellyn & Hobson, 2015). Cortical episodic memory networks then consist of sequential, associative pathways which converge at nodes/junctions which express patterned, probabilistic, experiential learning. The hidden units/nodes in neural networks may be analogous to these nodes/junctions in cortical episodic memory networks.

Given the authors' recognition, "that episodic memory could be the key to explaining human performance in naturalistic environments" it seems pertinent to expand their algorithms to encompass not only the logical, experiential, "step-by-step" learning patterns that predominate during wake but also the more complex, hyper-associative, experiential patterns identified during REM sleep. Admittedly, this extension may be, somewhat, futuristic. But research has already mimicked slow wave sleep which restored stability to neural networks engaged in unsupervised learning (Watkins, Kim, Sornborger, & Kenyon, 2020). Also, given the adaptivity, speed via massive parallelism, fault tolerance and optimality of neural networks (Bezdek, 1992) and that machine learning programmes trained on probabilistic reasoning are superior to the human brain for visual pattern recognition (Pavlus, 2016) it may be possible that, in the recurrent neural networks mobilized by the authors, REM sleep is not required for complex pattern recognition.

Experiences in natural environments may never be re-enacted in their entirety but they are certainly non-random and do not exclude expectations. Across evolutionary time, many interactions with the environment held dangers because of predators and competitors. To try to avoid dangers, humans needed to identify the probabilistic, behavioural patterns of predators and competitors. Such patterns would only be revealed over several episodes, as humans observed the associations that drove predator behaviour. For example, lions tend to visit waterholes at night when prey are abundant (lion presence is associated with nighttime) but, in the dry season, lions get so thirsty that they may be at the waterhole during the day (in the dry season, lion presence can be associated with daytime). Also, elephants tend to drive lions away from the waterhole, so the presence of elephants offers some safety (elephant presence is associated with lion absence). When many associations are at stake and/or actions must be fast, probabilistic associative patterns, derived from multiple episodes, drive expectations and learning during future interactions in the environment.

Much, probably the majority, of learning occurs unconsciously. Contemporarily, dangers may differ but unconscious expectations, formed through learning and retained as unconscious memories, still drive fast responses to threats (Öhman, Carlsson, Lundqvist, & Ingvar, 2007). Concomitantly, it would be expected that the elaborative, complex, associative, probabilistic and meaningful experiential patterns formed in REM sleep would be unconscious. Arzi et al. (2012) showed that new associations, learned during REM sleep, were retained as unconscious memories, then functioning, during wake, as unconscious expectations which drove actions.

Does the human unconscious have any parallels in machine learning? We know that machine learning algorithms inherit unconscious human biases, indeed they can amplify them (Thomas, 2018). Moreover, humans can continue to reproduce machine learning bias, even after they no longer use the algorithm (Vicente & Matute, 2023). Clearly, unconscious learning and memory are not subject to voluntary control. Although unconscious biases can be detrimental (e.g., to marginalized groups) inductive, unconscious (or implicit) biases are essential for faster information processing both in humans and machine learning. Inductive biases depend on the associations formed through experience, in machine learning such associations occur during unsupervised learning. Specifically, in the authors' meta-learned models, associations will be experiential.

In sum, the authors assert that their meta-learned models are "invaluable tools to study, analyse and understand the human mind." In humans, experiential, associative learning becomes associative memories that, then, drive expectations in subsequent learning. In the memory network architecture of the brain/mind these associative memories may take the form of the serial, associative pathways which underlie conscious learning. Equally, where these serial pathways cross at nodes/junctions, complex, hyper-associative, experiential patterns arise from elements of the different intersecting pathways. These patterns are retained as unconscious memories which associate elements of different experiences.

Association is fundamental to both learning and memory, whether conscious or unconscious. The authors' first research question of how people improve their learning abilities over time can be better addressed through acknowledging, first, the inextricability of learning and memory and, second, the role of conscious and unconscious association in each.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. None.

References

Arzi, A., Shedlesky, L., Ben-Shaul, M., Nasser, K., Oksenberg, A., Hairston, I. S., & Sobel, N. (2012). Humans can learn new information during sleep. *Nature neuroscience*, *15* (10), 1460–1465. doi: 10.1038/nn.3193

Bezdek, J. C. (1992). On the relationship between neural networks, pattern recognition and intelligence. *International Journal of Approximate Reasoning*, *6*(2), 85–107.

Brown, S. C., & Craik, F. I. M. (2000). Encoding and retrieval of information. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 93–107). Oxford University Press.

Davachi, L., & Wagner, A. D. (2002). Hippocampal contributions to episodic encoding: Insights from relational and item-based learning. *Journal of Neurophysiology*, *88*(2), 982–990. doi: 10.1152/jn.2002.88.2.982

Foer, J. (2011). *Moonwalking with Einstein: The art and science of remembering everything*. Allen Lane.

Fuster, J. M. (1999). *Memory in the cerebral cortex: An empirical approach to neural networks in the human and nonhuman primate*. MIT Press.

Llewellyn, S. (2013). Such stuff as dreams are made on? Elaborative encoding, the ancient art of memory, and the hippocampus. *Behavioral and Brain Sciences*, *36*(6), 589–607. doi: 10.1017/S0140525X12003135

Llewellyn, S., & Hobson, J. A. (2015). Not only… but also: REM sleep creates and NREM stage 2 instantiates landmark junctions in cortical memory networks. *Neurobiology of Learning and Memory*, *122*, 69–87. doi: 10.1016/j.nlm.2015.04.005

Öhman, A., Carlsson, K., Lundqvist, D., & Ingvar, M. (2007). On the unconscious subcortical origin of human fear. *Physiology & behavior*, *92*(1–2), 180–185. doi: 10.1016/j.physbeh.2007.05.057

Pavlus, J. (2016). Computers now recognize patterns better than humans can. *Scientific American*. Originally published as, "8. Sight-Reading Software" in Scientific American Magazine Vol. 315 No. 6 (December 2016), p. 41. doi: 10.1038/scientificamerican1216-39b

Thakral, P. P., Benoit, R. G., & Schacter, D. L. (2017). Characterizing the role of the hippocampus during episodic simulation and encoding. *Hippocampus*, *27*(12), 1275–1284. doi: 10.1002/hipo.22796

Thomas, R. (2018). Analyzing & preventing unconscious bias in machine learning. Retrieved from: QCon.ai 2018.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352–373.

Vicente, L., & Matute, H. (2023). Humans inherit artificial intelligence biases. *Scientific Reports*, *13*(1), 15737. doi: 10.1038/s41598-023-42384-8

Watkins, Y., Kim, E., Sornborger, A., & Kenyon, G. T. (2020). Using sinusoidally-modulated noise as a surrogate for slow-wave sleep to accomplish stable unsupervised dictionary learning in a spike-based sparse coding model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 360–361).

Yates, F. A. (1966). *The art of memory*. Routledge and Kegan Paul.

Meta-learning: Bayesian or quantum?

Antonio Mastrogiorgio*

Department of Psychological and Social Sciences, John Cabot University, Rome, Italy
www.johncabot.edu
mastrogiorgio.antonio@gmail.com
https://sites.google.com/site/mastrogiorgioantonio/

*Corresponding author.

doi:10.1017/S0140525X24000220, e154

Abstract

Abundant experimental evidence illustrates violations of Bayesian models across various cognitive processes. Quantum cognition capitalizes on the limitations of Bayesian models, providing a compelling alternative. We suggest that a generalized quantum approach in meta-learning is simultaneously more robust and flexible, as it retains all the advantages of the Bayesian framework while avoiding its limitations.

The use of the Bayesian framework in meta-learning represents an elegant way to bypass the strictures of an ex-ante specification of models of cognition. As proposed by Binz et al. (hereafter the Authors), Bayesian inference, building upon unconstrained interactions with the environment, represents a viable alternative to more traditional hand-designed learning algorithms.

However, Bayesian models come with inherent limitations, which researchers are rarely aware of. While scholars normally endorse Bayesian models because of their unconstrained features, they rarely consider that such models are actually "constrained" to the Kolmogorovian assumptions of classical probability theory.

Abundant experimental evidence illustrates violations of Bayesian models across various cognitive processes, including probability judgment errors, memory recognition, semantic spaces, information processing, learning, concept combination, and perception (e.g., Pothos and Busemeyer, 2022). These violations stem from the fact that many cognitive phenomena do not adhere to the law of total probability, along with the distributivity axiom, assumed in classical Kolmogorovian probability. Consequently, they do not admit a Bayesian operationalization.

Let us consider a stylized meta-learning process (similar to those discussed by the Authors) that can be operationalized through a Bayesian model. Suppose we hypothesize that learning performance can assume the (mutually exclusive) states P1 or P2, conditioned on meta-experience, which can assume the (mutually exclusive) states E1 or E2. Let's aim to predict the total probability, p, of the state P2. Consistent with a Bayesian framework, we can consider two cases: One, which we'll refer to as the "unconditioned case," where we only observe p(P2); and the other, referred to as the "conditioned case," where we observe the conditioned probabilities, p(P2|E1) and p(P2|E2).

Now, suppose that in experimental settings, the observed data reveal that in the "unconditioned case," the probability is p(P2) = 0.29, while in the "conditioned case," the probability is p(E1)·p (P2|E1) + p(E2)·p(P2|E2) = 0.59, where p(P2|E2) = 0.63.

We quickly realize that these experimental results are incompatible with a Bayesian framework: The total probability of P2 in the "unconditioned case" is inconsistent with that of the "conditioned case" (0.29 vs. 0.59), and the total probability of the "unconditioned case" is also lower than that in the "conditioned case," restricted to E2 (0.29 vs. 0.63). When evidence violates the law of total probability, Bayesian models reveal inadequacies (for further discussion, see Busemeyer & Wang, 2015; Busemeyer, Wang, & Lambert-Mogiliansky, 2009).

Such situations are paradoxical: They are somehow implausible to the extent that they do not admit a Bayesian formalization, yet they result from experimental evidence. When it comes to understanding such evidence, quantum cognition comes to the fore as a burgeoning field of research that dialectically capitalizes on the limitations of Bayesian models, providing a compelling alternative. A fundamental difference between Bayesian and quantum models – relevant for the present critique – lies in the fact that quantum models can account for such evidence precisely because they do not adhere to the law of total probability (for a comprehensive comparison between Bayesian and quantum models in cognition, refer to Bruza, Wang, & Busemeyer, 2015).

Continuing with the example discussed above, what differs between the "unconditioned case" and "conditioned case" is that, respectively, the non-observation or the observation of the conditioning variable (E) is not neutral for the final probability. In other words, the two models are incompatible and do not admit a mutually consistent formalization.

In a quantum framework, the violations of the total law of probability are due to *interference effects*, which occur when the conditioning variables are not observed. In our example, the interference between the two conditions (E1 and E2) implies that they behave like waves, where the interference can be either destructive (canceling) or constructive (resonating), affecting the final probability p(P2). On the contrary, when E is observed, the result is compatible with a classical framework as the act of measuring the mutually exclusive states of experience eliminates their interference (for an overview on the role of interference effects, see Busemeyer & Bruza, 2012).

The Authors wisely avoid claiming that meta-learning is the ultimate solution to every modeling problem, and they contemplate, in what they call "Intricate training processes," the possibility that "the resulting model [of meta-learning] does not fit the observed data."

We think that such situations are not just due, as implicitly suggested by the Authors, to the complexity of the scenarios, but to the fact that the tacit probabilistic assumptions of Bayesian models are somehow too restrictive.

More in detail, the plausibility of neurocognitive Bayesian foundations of meta-learning would require stronger justifications. Indeed, assuming that prefrontal circuits may constitute a meta-reinforcement learning system (cf., Wang et al., 2018) or, in general terms, that the brain is a Bayesian machine, matching top-down prediction with bottom-up experience (cf., Friston, 2010), would also imply assuming that Kolmogorovian probability ascribes to the biological realm.

The employment of Bayesian models is institutionalized to such an extent that their foundational assumptions are rarely contested in scientific debates. However, there are situations, like the one discussed here, in which Bayesian models reveal their limitations: Despite aspiring to provide a generalized framework for meta-learning, they inherently harbor very restrictive assumptions in probability theory.

On the contrary, quantum models exhibit greater flexibility, are more robust, and can offer a more sophisticated view of the neurocognitive mechanisms involved in human learning (cf., Mastrogiorgio, 2022).

However, quantum cognition is not a tout court alternative to Bayesian models but rather a generalization applicable to cases where the law of total probability is violated. This implies that Bayesian models represent a special case within the broader quantum framework: Quantum models reduce to Bayesian models when experimental evidence aligns with the requirements of the distributivity axiom and the law of total probability.

Precisely because we support the Authors' proposal of employing unconstrained logics in meta-learning, we also believe that a generalized quantum approach is simultaneously more robust and flexible, as it retains all the advantages of the Bayesian framework while avoiding its limitations.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

Bruza, P. D., Wang, Z., & Busemeyer, J. R. (2015). Quantum cognition: A new theoretical approach to psychology. *Trends in Cognitive Sciences*, *19*, 383–393.

Busemeyer, J. R., & Bruza, P. D. (2012). *Quantum models of cognition and decision*. Cambridge University Press.

Busemeyer, J. R., & Wang, Z. (2015). What is quantum cognition, and how is it applied to psychology?. *Current Directions in Psychological Science*, *24*(3), 163–169.

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

Busemeyer, J. R., Wang, Z., & Lambert-Mogiliansky, A. (2009). Empirical comparison of Markov and quantum models of decision making. *Journal of Mathematical Psychology*, *53*, 423–433.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Mastrogiorgio, A. (2022). A quantum predictive brain: Complementarity between Top-down predictions and bottom-up evidence. *Frontiers in Psychology*, *13*, 869894.

Pothos, E.M., & Busemeyer, J.R. (2022). Quantum cognition. *Annual Review of Psychology*, *73*, 749–778.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z.,… Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(6), 860–868.

Meta-learning as a bridge between neural networks and symbolic Bayesian models

R. Thomas McCoy[a] and Thomas L. Griffiths[b]*

[a]Department of Linguistics, Yale University, New Haven, CT, USA and
[b]Departments of Psychology and Computer Science, Princeton University, Princeton, NJ, USA
tom.mccoy@yale.edu
tomg@princeton.edu
https://rtmccoy.com/
http://cocosci.princeton.edu/tom/

*Corresponding author.

doi:10.1017/S0140525X24000116, e155

Abstract

Meta-learning is even more broadly relevant to the study of inductive biases than Binz et al. suggest: Its implications go beyond the extensions to rational analysis that they discuss. One noteworthy example is that meta-learning can act as a bridge between the vector representations of neural networks and the symbolic hypothesis spaces used in many Bayesian models.

Like many aspects of cognition, learning can be analyzed at multiple levels. At a high level (Marr's [1982] "computational" level) we can model learning by providing an abstract characterization of the learner's inductive biases: The preferences that the learner has for some types of generalizations over others (Mitchell, 1997). At a lower level, learning can be modeled by specifying the particular algorithms and representations that the learner uses to realize its inductive biases. For each of these levels, there are modeling traditions that have been successful: Rational analysis and Bayesian models are defined at the computational level, while neural networks are defined at the level of algorithm and representation. But how can we connect these different traditions? How can we work toward unified theories that bridge the divide between levels? In this piece, we agree with, and extend, Binz et al.'s point that meta-learning is a powerful tool for studying inductive biases in a way that spans levels of analysis.

Binz et al. describe how an agent can use meta-learning to derive inductive biases from its environment. This makes meta-learning well-suited for modeling situations where human inductive biases align with some problem that humans face – the situations that are well-covered by the paradigm of rational analysis (Anderson, 1990). As Binz et al. discuss, meta-learning can therefore be used to enable an algorithmically defined model (such as a neural network) to find the solution predicted by rational analysis, a procedure that bridges the divide between abstract rational solutions and specific algorithmic instantiations.

This direction laid out by Binz et al. is exciting. We argue that it can in fact be viewed as one special case within a broader space of possible lines of inquiry about inductive biases that meta-learning opens up. In the more general case, the Bayesian perspective allows us to define an inductive bias as a probability distribution over hypotheses. A neural network can meta-learn from data sampled from this distribution, giving it the inductive bias in question. The distribution that is used could be drawn from (an approximation of) a human's experience, in which case this framing matches the extension of rational analysis that Binz et al. advocate for. But it is also possible to use other approaches for defining this distribution, which can correspond to any probabilistic model. Since we can control probabilistic models, using a probabilistic model to define the distribution makes it possible to control the inductive biases that the meta-learned model ends up with (Lake, 2019; Lake & Baroni, 2023; McCoy, Grant, Smolensky, Griffiths, & Linzen, 2020). This allows us to take an inductive bias defined at Marr's computational level and distill it into a neural network defined at the level of algorithm and representation.

Traditionally, certain types of inductive biases have been associated with certain types of algorithms and representations: The strong inductive biases of Bayesian models have generally been based on discrete, symbolic representations (e.g., Goodman, Tenenbaum, Feldman, & Griffiths, 2008), while neural networks use continuous vector representations (Hinton, McClelland, & Rumelhart, 1986) and have weak inductive biases. However, meta-learning enables us to separately manipulate inductive biases and representations, making it possible to model previously inaccessible combinations of representations and inductive biases. One noteworthy example is that we can use meta-learning to give symbolic inductive biases to a neural network, allowing us to study whether and how structured hypothesis spaces (of the sort often used in Bayesian models) can be realized in a system with continuous vector representations (the type of representation that is central in both biological and artificial neural networks). Thus, while Binz et al. note that meta-learning can be used as an alternative to Bayesian models, another use of meta-learning is in fact to expand the applicability of Bayesian approaches by reconciling them with connectionist models – thereby bringing together two successful research traditions that have often been framed as antagonistic (e.g., Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; McClelland et al., 2010).

In our prior work, we have demonstrated the efficacy of this approach in the domain of language (McCoy & Griffiths, 2023). We started with a Bayesian model created by Yang and Piantadosi (2022), whose inductive bias is defined using a symbolic grammar. We then used meta-learning (specially, MAML: Finn, Abbeel, & Levine, 2017; Grant, Finn, Levine, Darrell, & Griffiths, 2018) to distill this Bayesian model's prior into a neural network. The resulting system had strong inductive biases of the sort traditionally found only in symbolic models, enabling this system to learn formal linguistic patterns from small numbers of examples despite being a neural network, a class of systems that normally requires far more examples to learn such patterns. Additionally, the flexible neural implementation of this system

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

made it possible to train it on naturalistic textual data, something that is intractable with the Bayesian model that we built on. Thus, meta-learning enabled the creation of a model that combined the complementary strengths of Bayesian and connectionist models of language learning.

These results show that inductive biases traditionally defined using symbolic Bayesian models can instead be realized inside a neural network. Therefore, symbolic inductive biases do not necessarily require inherently symbolic representations or algorithms. This demonstration provides one already-realized example of how meta-learning can advance our understanding of foundational questions about how different levels of cognition relate to each other, in ways that go beyond the realm of rational analysis.

Financial support. This material is based upon work supported by the National Science Foundation SBE Postdoctoral Research Fellowship under Grant No. 2204152 and the Office of Naval Research under Grant No. N00014-18-1-2873.

Competing interests. None.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.
Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 1126–1135.
Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.
Grant, E., Finn, C., Levine, S., Darrell, T., & Griffiths, T. (2018). Recasting gradient-based meta-learning as hierarchical Bayes. *International Conference on Learning Representations.*
Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.
Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 1. Foundations, pp. 77–109.
Lake, B. M. (2019). Compositional generalization through meta sequence-to-sequence learning. *Advances in Neural Information Processing Systems*, *32*.
Lake, B. M., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, *623*(7985), 115–121.
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman.
McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348–356.
McCoy, R. T., Grant, E., Smolensky, P., Griffiths, T. L., & Linzen, T. (2020). Universal linguistic inductive biases via meta-learning. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 737–743.
McCoy, R. T., & Griffiths, T. L. (2023). Modeling rapid language learning by distilling Bayesian priors into artificial neural networks. *arXiv preprint arXiv:2305.14701*.
Mitchell, T. M. (1997). *Machine learning*. McGraw Hill.
Yang, Y., & Piantadosi, S. T. (2022). One model for the learning of language. *Proceedings of the National Academy of Sciences*, *119*(5), e2021865119.

Meta-learned models beyond and beneath the cognitive

Mihnea Moldoveanu*

Desautels Centre for Integrative Thinking, Rotman School of Management, University of Toronto, Toronto, ON, Canada
mihnea.moldoveanu@rotman.utoronto.ca

*Corresponding author.

doi:10.1017/S0140525X24000141, e156

Abstract

I propose that meta-learned models, and in particular the situation-aware deployment of "learning-to-infer" modules can be advantageously extended to domains commonly thought to lie outside the cognitive, such as motivations and preferences on one hand, and the effectuation of micro- and coping-type behaviors.

An account of the ways in which locally meta-learned models can address difficulties arising from computational intractability of Bayesian inference in non-small worlds and difficulties in articulating inference problems over unknown state spaces have is overdue. We have for some time been aware that the experimental evidence base of behavioral decision theory admits of alternative plausible explanations via models of the experimental condition (meant to illustrate "base rate neglect," or availability of a plausible simile) that explain how the subject can seem biased to an observer while in fact engaging in a reasonable and ecologically successful pattern of inference (Moldoveanu & Langer, 2002; Marsh, Todd, & Gigerenzer, 2004). At the same time, incorporating the informational (say, "bits") and computational ("operations performed upon bits") costs of inferential calculations in models of cognition not only "makes sense" – as storage and calculation both require work – but can rationalize patterns of behavior that previously appeared to some as irrational or sub-rational (Gershman, Horvitz, & Tennenbaum, 2015; Moldoveanu, 2011).

The meta-learning program of inquiry can be extended to phenomena and episodes that lie beyond or on the fringes of what we would call "cognitive," at both the "upstream" (motives, motivations, identities) and "downstream" (motor behaviors, perceptual inferences) ends. Motivations, "preferences," and meta-preferences can be understood as the outcomes of a process by which one learns from one's own reactions to an environment's responses to one's own behaviors. At the other end, in-the-moment "heedful coping" for the purpose of getting a "maximal grip" (Merleau Ponty, 2012[1974]) on an immediate physical or interpersonal environment can be described in ways that make transparent the structure of the inferential learning problem and the function that one is trying to approximate, for example: Producing situation-specific, successful combinations of vertical and horizontal pressures through one's digits in order to balance a large, flat object, or of activations of muscles to produce particular facial and postural expressions meant to cause someone else to do, say or think in a particular way.

Consider, first, the "motivational" nexus of motivation-identity-preference. The idea that one infers one's own "attitudes" from one's behavior in situations whose affordances encode "choicefulness" has been around for a while (Bem, 1967), but the problem of "learning to prefer (X to Y, say)" is neither immediately self-evident or "given" nor, once posed, computationally simple (or, even tractable). But, once preferences are understood as dispositions to act in particular ways or combinations of ways in a context and encoded by conditional probabilities linking combinations of sensorial fields and internal states to actions, the self-referential problem of inferring "What do I like?/What motivates me?" from observations of "what do I do (or, choose to do) in situations in which…?" can get off the ground. One can develop preferences that are highly specific (combinations of ingredients mixed in precise sequences and proportions to make a "dish") or general ("I prefer injustice to disorder") and motivations that are domain-specific ("decentralized decision rights as an approach to managing *this* team on *this* task") or less so ("to enable and facilitate other people's sense of autonomy"). Preferences and motivations can be more or less sophisticated (in terms of the number of features of a "situation" the inference of preference depends on, and the logical depth or computational complexity of the inference algorithm) and more or less context-adaptive, and more or less susceptible to recursive refinement upon experiencing the ways in which one behaves in different environments. Thus, being able to adaptively modify the informational and computational complexity of the learning algorithm adds a much-needed degree of freedom to the modelers toolkit. Both preferences and motivations may be learned without an explicit awareness of that which is learned or that which learning is conditioned by. "Learning to prefer" (or learning to be the self one, motivationally, is) can thus be phrased in a way that tracks "learning to infer", provided we can successfully formulate a local objective or cost function the inferential process meliorates or seeks to extremize. "Internal conflicts" appear as neither pathological nor irrational: An optimal (or ecologically adapted in virtue of its adaptiveness) brain can comprise a set of independent agents that have conflicting objectives (Livnat & Pippenger, 2006) but are jointly adaptive to environmental niches its organism copes with frequently enough.

Second, consider the "in-the-moment" perception-sensation-effectuation nexus, which has to do with trying to get a target variable to take on a certain value within a certain time window using the least or a fixed amount of energy – such as controlling the vertical displacement of a tray of containers containing hot liquids, the horizontal displacement of an inverted pendulum, the angular velocity, height, and vertical velocity of a coin used in a coin toss or even a dynamical network of physiognomical micro-responses in an emotionally charged meeting with several attendees. In such cases, the inverse, counterfactual-dependent problem of "causal inference" is replaced by the forward, direct problem of effectuation (or control): One is attempting to produce and in some cases also maintain a specific set of values of variables X_T (vertical displacement of a tray of liquids, perceived visceral or emotional responses of another person) that form part of the state space $\boldsymbol{X}$ of a dynamically evolving system $\boldsymbol{X_t = F(X,U,\ t)}$ by making specific changes to one or more "input" or lever variables $\boldsymbol{U}$ on the basis of observations of some proxy variables $\boldsymbol{Y}$ that encode or register filtered, biased states of $\boldsymbol{X}$ via $\boldsymbol{Y = G(X,t)}$. In this case, the underlying inference problem can be posed in terms of maximizing the time-bounded and energy-efficient controllability and observability of $\boldsymbol{X_t = F(X,U,\ t);\ Y = G(X,t)}$ for instance by the optimal choice and placement of the "lever" or "driver" nodes (Li and Laszlo Barabassi, 2016), or in terms of making changes to the structural properties of $\boldsymbol{X_t = F(X,U,\ t);\ Y = G(X,t)}$ in ways that alter its temporal dynamics or time constants (for instance, learning to control a tremor by using different combinations or muscle groups for effecting a fine movement, which changes the parameters of $\boldsymbol{X_t = F(X,U,\ t);\ Y = G(X,t)}$ and thus the pole-zero distribution of its transfer function).

Financial support. This research was funded by the Desautels Centre for Integrative Thinking, Rotman School of Management, University of Toronto.

Competing interest. None.

References

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, *74*, 183–200.

Gershman, S. J., Horvitz, E. J., & Tennenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds and machines. *Science*, *349* (6245), 273–278.

Liu, Y. -Y., & Laszlo-Barabassi, A. (2016). Control principles of complex systems. *Review of Modern Physics*, *88*(3), 1–58.

Livnat, A., & Pippenger, N. (2006). An optimal brain can be composed of conflicting agents. *Proceedings of the National Academy of Sciences*, *103*(9), 3198–3202.

Marsh, B., Todd P. M., & Gigerenzer, G. (2004). Cognitive heuristics: Reasoning the fast and frugal way. In J. P. Leighton, & R. J. Sternberg (Eds.), *The nature of reasoning* (pp. 273–287). Cambridge University Press.

Merleau Ponty, M. (2012)(1974). *The phenomenology of perception*. Routledge.

Moldoveanu, M. C. (2011). *Inside man: The discipline of modeling human ways of being*. Stanford University Press.

Moldoveanu, M. C., & Langer, E. J. (2002). False memories of the future: A critique of the applications of probabilistic reasoning to the study of cognitive processes. *Psychological Review*, *109*(2), 358–375.

Meta-learned models as tools to test theories of cognitive development

Kate Nussenbaum[a]* and Catherine A. Hartley[b]*

[a]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA and [b]Department of Psychology, New York University, New York, NY, USA
katenuss@princeton.edu
cate@nyu.edu
https://www.katenuss.com/
https://www.hartleylab.org/

*Corresponding authors.

doi:10.1017/S0140525X24000281, e157

Abstract

Binz et al. argue that meta-learned models are essential tools for understanding adult cognition. Here, we propose that these models are particularly useful for testing hypotheses about why learning processes change across development. By leveraging their ability to discover optimal algorithms and account for capacity limitations, researchers can use these models to test competing theories of developmental change in learning.

Binz et al. argue compellingly for meta-learning as a tool to understand adult cognition, but their vision is incomplete: Meta-learned models are particularly apt tools for studying development. As the authors note, meta-learned models provide a natural foundation for theorizing about learning to learn (Wang, 2021) and for understanding experience-driven changes in goal-directed behavior (Nussenbaum & Hartley, in press). Here, however, we consider the authors' central argument and focus not on meta-learning algorithms or the process of "learning to learn," but rather on meta-learned models as a *tool* for understanding developmental changes in learning. To date, research has revealed age-related shifts in the algorithms and neural circuitry that underlie learning (Bolenz, Reiter, & Eppinger, 2017; Gualtieri & Finn, 2022; Hartley, Nussenbaum, & Cohen, 2021; Nussenbaum & Hartley, 2019; Raab & Hartley, 2018), but understanding why these shifts occur has proven more difficult.

Disentangling whether age-related changes in learning reflect adaptation to age-varying "external" ecological problems *or* "internal" changes in cognitive capacity has been difficult, in part because developmentalists have lacked the formalizations needed to make specific predictions about how these two factors might drive age-related changes in behavior. As an example, in many studies of reinforcement learning, participants are tasked with earning the most reward (e.g., points or money) by selecting between different options. In these tasks, children tend to make "noisier" or more exploratory choices, which typically result in poorer performance (Nussenbaum & Hartley, 2019) but enhanced learning of task structure (Blanco & Sloutsky, 2021; Liquin & Gopnik, 2022; Sumner et al., 2019). These findings have led researchers to theorize that children and adults have optimized their learning computations to maximize reward over different types of environments – children's learning contexts may have features (e.g., greater reward stochasticity, longer temporal horizons) that favor exploration over immediate reward gain (Gopnik, 2020). An alternative account, however, is that children's "noisier" behavior reflects a more limited cognitive capacity (Craik & Bialystok, 2006; Ruel, Devine, & Eppinger, 2021), which constrains their "optimization" of learning computations.

By enabling separable manipulation of external experience and internal cognitive capacity, meta-learned models may help arbitrate between these accounts of developmental change. As Binz et al. note, through meta-learning, the optimal algorithm emerges through experience, and is shaped by the distribution of environments on which the model is trained. Thus, these models can be used to test how differences in "training" experience might yield different patterns of learning at "test." Examining how training environments influence model behavior within the same tasks that have been used to study cognitive development can provide insight into the types of environments for which children, adolescents, and adults have optimized their learning computations. As one example, Binz and Schulz (2022) found that increasing the variance in the rewards that a model experienced during training led to exploration decisions at test that more closely approximated those of human learners, suggesting that people may have tuned their learning computations for environments with more stochasticity than the task context. Importantly, meta-learned models can provide insight into how optimal learning is shaped by exposure to highly complex environments in which multiple features vary and interact – environments for which analytically derived "solutions" are intractable but that are more reflective of real-world contexts than the simple tasks that have been used in most prior research.

At the same time, meta-learned models can account for how changes in capacity limitations may yield developmental changes in learning. As Binz et al. note, the complexity of meta-learned models can be manipulated easily, particularly when they are implemented as neural networks. Manipulating capacity constraints and comparing model behavior to that of learners at different ages can thus provide insight into whether developmental changes in learning are well accounted for by theories of "resource-rationality." Binz and Schulz (2022) found, for example, that increasing the complexity of the algorithms that a meta-learning network could implement led to changes in patterns of directed exploration that mirrored those that occur across adolescence (Somerville et al., 2017). This result exemplifies how, rather than emerging from differences in external "training" experience, age-related changes in learning can also be driven by differences in cognitive capacity.

Further, neural network implementations of meta-learning enable separable manipulations of "algorithmic complexity" via network weights and "computational complexity" via network activations. Future developmental modeling work could explore the ramifications of constraints on algorithmic *versus* computational complexity. Numerous studies have revealed a dissociation between children's knowledge of rules or structure and their ability to leverage them to guide behavior (Decker, Otto, Daw, & Hartley, 2016; Zelazo, Frye, & Rapus, 1996). The types of structural knowledge that can be acquired at different ages may depend on algorithmic complexity, while use of that knowledge may rely on processes like working memory and proactive cognitive control, which may instantiated via complex computations. Behavioral dissociations between algorithmic and computational complexity may be mirrored by neural dissociations. Age-related change in brain structure or the wiring of neural circuitry may be analogous to age-related change in network weights and relate more strongly to algorithmic complexity, whereas age-related change in patterns of neural activation during learning may be more closely related to the network activations that underlie computational complexity. Thus, predictions about how age-related change in these two forms of capacity limitations affect learning could be further tested and constrained with neural data.

While we have suggested that researchers can leverage meta-learned models to more explicitly test whether children optimize their behavior for different environments *or* with different constraints, this dichotomy is likely false. Experience and constraints interact throughout the lifespan, and the changing "constraints" implemented by neurobiology may themselves serve an adaptive function – it may be the case that the complexity of the learning algorithms an organism can implement and execute systematically increases *through* exposure to increasingly varied and complex environments. Meta-learned models of cognition thus have the potential to address questions of longstanding interest in developmental science, while empirical developmental research provides a valuable testbed for the theoretical utility of these computational tools.

Acknowledgements. We thank Rheza Budiono for helpful feedback.

Financial support. Preparation of this commentary was supported by a CV Starr Foundation Fellowship (to K. N.).

Competing interest. None.

References

Binz, M., & Schulz, E. (2022). Modeling human exploration through resource-rational reinforcement learning. *Advances in Neural Information Processing Systems*, *35*, 31755–31768.

Blanco, N. J., & Sloutsky, V. M. (2021). Systematic exploration and uncertainty dominate young children's choices. *Developmental Science*, *24*(2), e13026.

Bolenz, F., Reiter, A. M. F., & Eppinger, B. (2017). Developmental changes in learning: Computational mechanisms and social influences. *Frontiers in Psychology*, *8*, 2048.

Craik, F. I. M., & Bialystok, E. (2006). Cognition through the lifespan: Mechanisms of change. *Trends in Cognitive Sciences*, *10*(3), 131–138.

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychological Science*, *27*(6), 848–858.

Gopnik, A. (2020). Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *375*(1803), 20190502.

Gualtieri, S., & Finn, A. S. (2022). The sweet spot: When children's developing abilities, brains, and knowledge make them better learners than adults. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *17*(5), 1322–1338.

Hartley, C. A., Nussenbaum, K., & Cohen, A. O. (2021). Interactive development of adaptive learning and memory. *Annual Review of Developmental Psychology*, *3*, 59–85.

Liquin, E. G., & Gopnik, A. (2022). Children are more exploratory and learn more than adults in an approach-avoid task. *Cognition*, *218*, 104940.

Nussenbaum, K., & Hartley, C. A. (2019). Reinforcement learning across development: What insights can we draw from a decade of research? *Developmental Cognitive Neuroscience*, *40*, 100733.

Nussenbaum, K., & Hartley, C.A. (in press). Understanding the development of reward learning through the lens of meta-learning. *Nature Reviews Psychology*.

Raab, H. A., & Hartley, C. A. (2018). The development of goal-directed decision-making. In R. Morris, A. Bornstein, & A. Shenhav (Eds.), *Goal-directed decision making: Computations and neural circuits* (pp. 279–308). Elsevier Academic Press.

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

Ruel, A., Devine, S., & Eppinger, B. (2021). Resource-rational approach to meta-control problems across the lifespan. *Wiley Interdisciplinary Reviews. Cognitive Science*, *12*(5), e1556.

Somerville, L. H., Sasse, S. F., Garrad, M. C., Drysdale, A. T., Abi Akar, N., Insel, C., & Wilson, R. C. (2017). Charting the expansion of strategic exploratory behavior during adolescence. *Journal of Experimental Psychology. General*, *146*(2), 155–164.

Sumner, E., Li, A. X., Perfors, A., Hayes, B., Navarro, D., & Sarnecka, B. W. (2019). The exploration advantage: Children's instinct to explore allows them to find information that adults miss. https://doi.org/10.31234/osf.io/h437v

Wang, J. X. (2021). Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences*, *38*, 90–95.

Zelazo, P. D., Frye, D., & Rapus, T. (1996). An age-related dissociation between knowing rules and using them. *Cognitive Development*, *11*(1), 37–63.

Probabilistic programming versus meta-learning as models of cognition

Desmond C. Ong[a]*, Tan Zhi-Xuan[b], Joshua B. Tenenbaum[b] and Noah D. Goodman[c,d]

[a]Department of Psychology, University of Texas at Austin, Austin, TX, USA; [b]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA; [c]Department of Psychology, Stanford University, Stanford, CA, USA and [d]Department of Computer Science, Stanford University, Stanford, CA, USA
desmond.ong@utexas.edu
xuan@mit.edu
jbt@mit.edu
ngoodman@stanford.edu
https://cascoglab.psy.utexas.edu/desmond/
https://cocosci.mit.edu/
https://ztangent.github.io/
https://cocolab.stanford.edu/

*Corresponding author.

doi:10.1017/S0140525X24000153, e158

Abstract

We summarize the recent progress made by probabilistic programming as a unifying formalism for the probabilistic, symbolic, and data-driven aspects of human cognition. We highlight differences with meta-learning in flexibility, statistical assumptions and inferences about cogniton. We suggest that the meta-learning approach could be further strengthened by considering Connectionist *and* Bayesian approaches, rather than exclusively one or the other.

Connectionist-versus-Bayesian debates have occurred in cognitive science for decades (e.g., Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; McClelland et al., 2010), with each side progressing in theory, models, and algorithms, in turn impelling the other side to advance, resulting in a cycle of fruitful engagement. The recent summary of the meta-learning paradigm that Binz et al. proposed in the target article bridges the two by proposing how meta-learning in recurrent neural networks can address some of the traditional challenges of Bayesian approaches. But, by failing to recognize and engage with the latest iteration of Bayesian modeling approaches – including probabilistic programming as a unifying paradigm for probabilistic, symbolic, and differentiable computation (Cusumano-Towner, Saad, Lew, & Mansinghka, 2019) – this article fails to push the meta-learning paradigm as far as it could go.

The authors begin their defense of meta-learning by citing the intractability of *exact* Bayesian inference. However, this fails to address how and why meta-learning is superior to *approximate* inference for modeling cognition. As the authors themselves note, Bayesian modelers use a variety of approximate inference methods, including neural-network-powered variational inference (Dasgupta, Schulz, Tenenbaum, & Gershman, 2020; Kingma & Welling, 2013), Markov chain Monte Carlo (Ullman, Goodman, & Tenenbaum, 2012), and Sequential Monte Carlo methods (Levy, Reali, & Griffiths, 2008; Vul, Alvarez, Tenenbaum, & Black, 2009), which have all shown considerable success in modeling how humans perform inference (or fail to) in presumably intractable settings. As such, it is hardly an argument in favor of meta-learning – and against "traditional" Bayesian models – that *exact* inference is intractable.

This omission is just one way in which the article fails to engage with a modern incarnation of the Bayesian modeler's toolkit – Probabilistic Programming. In the past two decades, we have seen the development of probabilistic programming as unifying formalism for modeling the probabilistic, symbolic, and data-driven aspects of human cognition (Lake, Salakhutdinov, & Tenenbaum, 2015), as embodied in probabilistic programming language such as Church (Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2012), webPPL (Goodman & Stuhlmüller, electronic), Pyro (Bingham et al., 2019), and Gen (Cusumano-Towner et al., 2019). These languages enable modelers to explore a much wider range of computational architectures than the standard meta-learning setup, which requires modelers to reformulate human cognition as a sequence prediction problem. Probabilistic programming allows modelers to unite the strengths of general-purpose predictors (i.e., neural networks) with theoretically informed constraints and model-based reasoning. For instance, Ong, Soh, Zaki, and Goodman (2021) showed how reasoning about others' emotions can be modeled by combining the constraints implied by cognitive appraisal theory with bottom-up representations learnt via neural networks from emotional facial expressions. Similarly, several recent papers have shown how the linguistic abilities of large language models (LLMs) can be integrated with rational models of planning, communication, and inverse planning (Wong et al., 2023; Ying, Zhi-Xuan, Mansinghka, & Tenenbaum, 2023), modeling human inferences that LLM-based sequence prediction alone struggle with (Zhi-Xuan, Ying, Mansinghka, & Tenenbaum, 2024).

What flexibility does probabilistic programming afford over pure meta-learning? As the article notes, one potential benefit of meta-learning is that it avoids the need for a specific Bayesian model to perform inference over. Crucially, meta-learning achieves this by having access to sufficiently similar data at training and test time, such that the meta-learned algorithm is sufficiently well-adapted to the implied class of data-generating processes. Human cognition is much more adaptive. We do not simply adjust our learning to fit past distributions; we also construct, modify, abstract, and refactor entire *theories* about how the world works (Rule, Tenenbaum, & Piantadosi, 2020; Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Ullman & Tenenbaum, 2020), reasoning with such theories on downstream tasks (Tsividis et al., 2021). This capacity is not captured by pure meta-learning, which occurs "offline." By contrast, probabilistic programming allows modeling these patterns of thought: Theory building can be formulated as program induction (Lake et al., 2015; Saad, Cusumano-Towner, Schaechtle, Rinard, & Mansinghka, 2019), refactoring as program merging (Hwang, Stuhlmüller, & Goodman, 2011), and abstraction-guided reasoning as coarse-to-fine inference (Cusumano-Towner, Bichsel, Gehr, Vechev, & Mansinghka, 2018; Stuhlmüller, Hawkins, Siddharth, & Goodman, 2015). Inference meta-programs (Cusumano-Towner et al., 2019; Lew et al., 2023) allow us to model how people invoke modeling and inference strategies as needed: One can employ meta-learned inference when one believes a familiar model applies, but also flexibly compute inferences when a model is learned, extended, or abstracted. On this view, meta-learning has an important role to play in modeling human cognition, but not for all of our cognitive capacities.

Another way of understanding the relationship between meta-learning and probabilistic programming is that the former uses implicit statistical assumptions while the latter's assumptions are explicit. Meta-learning assumes that the structure of the world is conveyed in the statistical structure of data across independent instances. With sufficient coverage of the training distribution, flexible deep learning approaches fit this structure and use it to generalize. But they may not do so in a way that may provide any insight into the *computational* problem being solved by humans. Probabilistic programs, by contrast, explicitly hypothesize the statistical patterns to be found in data, providing constraints that, if satisfied, yield insights for cognition. This implicit–explicit distinction both frames the relative value of the approaches and suggests an alternative relation: A Bayesian model need not subsume or integrate what is learned by a deep learning model, but simply *explicate* it, at a higher level of analysis. Through this lens, having to specify an inference problem is not a limitation, but a virtue.

The best of both worlds will be to compose and further refine these paradigms, such as using deep amortized inference (like meta-learning for Probabilistic Programming), using Bayesian tools (and other tools for mechanistic interpretation) to understand the results of meta-learning, or constructing neurosymbolic models (e.g., by grounding the outputs of meta-learned models in probabilistic programs, as in Wong et al., 2023). As a very recent example, Zhou, Feinman, and Lake (2024) proposed a neurosymbolic program induction model to capture human visual learning, using both Bayesian program induction and meta-learning, achieving the best of both approaches: Interpretability and parsimony, as well as capturing additional variance using flexible function approximators. We believe that the field should move beyond "Connectionist-versus-Bayesian" debates to instead explore hybrid "Connectionist-*and*-Bayesian" approaches.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interest. None.

References

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., & … Goodman, N. D. (2019). Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research*, *20*(1), 973–978.

Cusumano-Towner, M., Bichsel, B., Gehr, T., Vechev, M., & Mansinghka, V. K. (2018). Incremental inference for probabilistic programs. In Proceedings of the 39th ACM SIGPLAN Conference on Programming Language Design and Implementation (pp. 571–585).

Cusumano-Towner, M. F., Saad, F. A., Lew, A., & Mansinghka, V. K. (2019). Gen: A general-purpose probabilistic programming system with programmable inference. In Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI '19).

Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review*, *127*(3), 412.

Goodman, N. D., Mansinghka, V., Roy, D. M., Bonawitz, K., & Tenenbaum, J. B. (2012). Church: a language for generative models. arXiv preprint arXiv:1206.3255.

Goodman N. D., & Stuhlmüller, A. (electronic). The design and implementation of probabilistic programming languages. Retrieved from http://dippl.org.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Hwang, I., Stuhlmüller, A., & Goodman, N. D. (2011). Inducing probabilistic programs by Bayesian program merging. arXiv preprint arXiv:1110.5667.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Levy, R., Reali, F., & Griffiths, T. (2008). Modeling the effects of memory on human online sentence processing with particle filters. *Advances in Neural Information Processing Systems*, *21*, 937–944.

Lew, A. K., Matheos, G., Zhi-Xuan, T., Ghavamizadeh, M., Gothoskar, N., Russell, S., & Mansinghka, V. K. (2023). SMCP3: Sequential Monte Carlo with probabilistic program proposals. In International Conference on Artificial Intelligence and Statistics (pp. 7061–7088). PMLR.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348–356.

Ong, D. C., Soh, H., Zaki, J., & Goodman, N. D. (2021). Applying probabilistic programming to affective computing. *IEEE Transactions on Affective Computing*, *12*(2), 306–317.

Rule, J. S., Tenenbaum, J. B., & Piantadosi, S. T. (2020). The child as hacker. *Trends in Cognitive Sciences*, *24*(11), 900–915.

Saad, F. A., Cusumano-Towner, M. F., Schaechtle, U., Rinard, M. C., & Mansinghka, V. K. (2019). Bayesian synthesis of probabilistic programs for automatic data modeling. *Proceedings of the ACM on Programming Languages*, *3*(POPL), 1–32.

Stuhlmüller, A., Hawkins, R. X., Siddharth, N., & Goodman, N. D. (2015). Coarse-to-fine sequential Monte Carlo for probabilistic programs. arXiv preprint arXiv:1509.02962.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Tsividis, P. A., Loula, J., Burga, J., Foss, N., Campero, A., Pouncy, T., & …Tenenbaum, J. B. (2021). Human-level reinforcement learning through theory-based modeling, exploration, and planning. arXiv preprint arXiv:2107.12544.

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*(4), 455–480.

Ullman, T. D., & Tenenbaum, J. B. (2020). Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, *2*, 533–558.

Vul, E., Alvarez, G., Tenenbaum, J., & Black, M. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. *Advances in Neural Information Processing Systems*, *22*, 1955–1963.

Wong, L., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., & Tenenbaum, J. B. (2023). From word models to world models: Translating from natural language to the probabilistic language of thought. arXiv preprint arXiv:2306.12672.

Ying, L., Zhi-Xuan, T., Mansinghka, V., & Tenenbaum, J. B. (2023). Inferring the goals of communicating agents from actions and instructions. In Proceedings of the AAAI Symposium Series (Vol. 2, No. 1, pp. 26–33).

Zhi-Xuan, T., Ying, L., Mansinghka, V., & Tenenbaum, J. B. (2024). Pragmatic instruction following and goal assistance via cooperative language guided inverse plan search. In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems.

Zhou, Y., Feinman, R., & Lake, B. M. (2024). Compositional diversity in visual concept learning. *Cognition*, *244*, 105711.

Meta-learning in active inference

O. Penacchio[a]* and A. Clemente[b]

[a]Computer Science Department, Autonomous University of Barcelona, and School of Psychology and Neuroscience, University of St Andrews, Barcelona, Spain and [b]Department of Cognitive Neuropsychology, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany
op5@st-andrews.ac.uk
ana.clemente@ae.mpg.de
https://openacchio.github.io/
https://www.aesthetics.mpg.de/institut/mitarbeiterinnen/ana-clemente.html

*Corresponding author.

doi:10.1017/S0140525X24000074, e159

Abstract

Binz et al. propose meta-learning as a promising avenue for modelling human cognition. They provide an in-depth reflection on the advantages of meta-learning over other computational models of cognition, including a sound discussion on how their proposal can accommodate neuroscientific insights. We argue that active inference presents similar computational advantages while offering greater mechanistic explanatory power and biological plausibility.

Binz et al. provide a meritorious survey of the prospects offered by meta-learning for building models of human cognition. Among the main assets of meta-learning discussed by Binz et al. is the capacity to learn inductive bias from experience independently of the constraints enforced by the modeller. Further, Binz et al. showcase the capacity of meta-learning algorithms to approximate Bayesian inference, the gold standard for modelling rational analysis. Finally, they claim that meta-learning offers an unequalled framework for constructing rational models of human cognition that incorporate insights from neuroscience. We propose that an alternative theory of cognition, active inference, shares the same strengths as Binz et al.'s proposal while establishing precise and empirically validated connections to neurobiological mechanisms underlying cognition.

Learning from experience has become a benchmark in all fields aiming to understand and emulate natural intelligence and might be the next driver of developments in artificial intelligence (Zador & Tsao, 2023). In this regard, meta-learning joins other frameworks with the potential to advance the understanding of human cognition as it allows learning algorithms to adapt to experience beyond the modeller's intervention. However, active inference provides a distinct advantage as it has the purpose of modelling and understanding how agents engage with their environment. In active inference, cognitive agents – or algorithms – learn through experience by continually refining their internal model of the environment or the task at hand (Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016).

Another critical aspect of Binz et al.'s proposal is their insistent reference to Bayes' optimality. The concept has well-grounded theoretical and empirical foundations (Clark, 2013) that make it a good standard, justifying Binz et al.'s eagerness to probe their approach against Bayes' optimality. Their algorithm approximates Bayes' optimality with the mathematical consequence that any cognitive phenomenon accounted for by Bayesian inference can, in theory, be accounted for by meta-learning. However, the flexibility of Binz et al.'s approach to meta-learning entails a reduced interpretability of the resulting models. In active inference, posterior distributions are inferred using the free-energy principle, a variational approach to Bayesian inference that also approximates intractable computations but within a fully interpretable architecture (see below).

We also commend Binz et al. for describing meta-learning's capacity to incorporate insights from neuroscience, a requisite for a computational understanding of cognition (Kriegeskorte & Douglas, 2018). Yet, the biologically inspired elements introduced are *ad hoc* and case-dependent, as explicitly stated in their conclusion. Their meta-learning models are not motivated by a fundamental biological principle but are conceived as a powerful tool to enhance learning. By contrast, active inference directly translates into neural mechanisms (Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2017) and originates from a single unifying principle: the imperative for organisms to avoid *surprising* states, implemented by a continuous loop between drawing hypotheses on hidden states (e.g., mean length of an insect species) and observations (e.g., length of a particular specimen) (Friston, 2010). This principle aligns with the Helmoltzian perspective of perception as inference and subsequent Bayesian brain theories. The variational inferential dynamic when receiving new *observations* can be naturally cast into a constant bidirectional *message passing* with direct neural implementation as ascending prediction errors and descending predictions (Pezzulo, Parr, & Friston, 2024). Importantly, these mechanisms are common to all active inference models and enjoy empirical support (e.g., Bastos et al., 2012; Schwartenbeck, FitzGerald, Mathys, Dolan, & Friston, 2015).

Learning is a central construct in active inference. Agents constantly update their generative models based on observations and prediction errors, with the imperative of reducing prediction errors. Generative models represent alternative hypotheses about task execution and associated outcomes (e.g., estimating the average length of a species). These hypotheses, and possibly all components of the generative models, are tested and refined through the agent's experience (Friston et al., 2016). This experience-dependent plasticity has two fundamental assets.

First, learning in active inference is directly and naturally interpreted in terms of biologically plausible neuronal mechanisms. The updates of all components of the generative models are driven by co-occurrences between predicted outcomes (in postsynaptic units in the neuronal interpretation sketched above) and (presynaptic) observational inputs in a process reminiscent of Hebbian learning (Friston et al., 2016). Consequently, active inference is cast as a process theory that can draw specific empirical predictions on neuronal dynamics (Whyte & Smith, 2021).

Second, and crucially, agents in active inference learn the reliability of inputs and prediction errors. This precision estimation is akin to learning meta-parameters (as per Binz et al.) as it entails a weighting process that prioritises reliable sources over uninformative inputs. The balance between exploration and exploitation (central constructs in cognition reflecting epistemic affordances and pragmatic value, respectively) rests upon mechanisms with direct neurobiological substrate in terms of dopamine release, with important implications for rational decision-making – for example, in two-armed bandit tasks (Schwartenbeck et al., 2019), maze navigation (Kaplan & Friston, 2018) and computational psychiatry (Smith, Badcock, & Friston, 2021). Another essential strength of active inference for implementing meta-learning is its natural hierarchical extension. Upper levels can control parameters of lower levels, enabling inference at different timescales whereby learning at lower levels is optimised over time by top-down adjustments from upper levels, which has direct neuronal interpretation in multi-scale hierarchical brain organisation (Pezzulo, Rigoli, & Friston, 2018).

Model preference depends on performance and, primarily, on the scientific question at hand. To understand cognition and its mechanistic underpinnings, models whose components and articulations can be directly interpreted in terms of neural mechanisms are essential. Active inference is a principled, biologically plausible and fully interpretable model of cognition with promising applications to artificial intelligence that accounts for neurobiological and psychological phenomena. We contend that it

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

provides a comprehensive model for understanding biological systems and improving artificial cognition.

Financial support. O. P. was funded by a Maria Zambrano Fellowship for the attraction of international talent for the requalification of the Spanish university system—Next Generation EU.

Competing interests. None.

References

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711. https://doi.org/10.1016/j.neuron.2012.10.038

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews*, *68*, 862–879. https://doi.org/10.1016/j.neubiorev.2016.06.022

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, *29*(1), 1–49. https://doi.org/10.1162/NECO_a_00912

Kaplan, R., & Friston, K. J. (2018). Planning and navigation as active inference. *Biological Cybernetics*, *112*(4), 323–343. https://doi.org/10.1007/s00422-018-0753-2

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148–1160. https://doi.org/10.1038/s41593-018-0210-5

Pezzulo, G., Parr, T., & Friston, K. (2024). Active inference as a theory of sentient behavior. *Biological Psychology*, *186*, 108741. https://doi.org/10.1016/j.biopsycho.2023.108741

Pezzulo, G., Rigoli, F., & Friston, K. J. (2018). Hierarchical active inference: A theory of motivated control. *Trends in Cognitive Sciences*, *22*(4), 294–306. https://doi.org/10.1016/j.tics.2018.01.009

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., & Friston, K. (2015). The dopaminergic midbrain encodes the expected certainty about desired outcomes. *Cerebral Cortex*, *25*(10), 3434–3445. https://doi.org/10.1093/cercor/bhu159

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H. B., Kronbichler, M., & Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *eLife*, *8*, e41703. https://doi.org/10.7554/eLife.41703

Smith, R., Badcock, P., & Friston, K. J. (2021). Recent advances in the application of predictive coding and active inference models within clinical neuroscience. *Psychiatry and Clinical Neurosciences*, *75*(1), 3–13. https://doi.org/10.1111/pcn.13138

Whyte, C. J., & Smith, R. (2021). The predictive global neuronal workspace: A formal active inference model of visual consciousness. *Progress in Neurobiology*, *199*, 101918. https://doi.org/10.1016/j.pneurobio.2020.101918

Zador, A., Escola, S., Richards, B., Ölveczky, B., Bengio, Y., Boahen, K., Tsao, D. (2023). Catalyzing next-generation Artificial Intelligence through NeuroAI. *Nature Communications*, *14*(1), 1597. https://doi.org/10.1038/s41467-023-37180-x

Quo vadis, planning?

Jacques Pesnot-Lerousseau[a,b] and
Christopher Summerfield[c]*

[a]Institute for Language, Communication, and the Brain, Aix-Marseille Univ, Marseille, France; [b]Aix Marseille Univ, Inserm, INS, Inst Neurosci Syst, Marseille, France and [c]Department of Experimental Psychology, University of Oxford, Oxford, UK
christopher.summerfield@psy.ox.ac.uk
jacques.pesnot-lerousseau@univ-amu.fr
https://humaninformationprocessing.com/

*Corresponding author.

doi:10.1017/S0140525X24000190, e160

Abstract

Deep meta-learning is the driving force behind advances in contemporary AI research, and a promising theory of flexible cognition in natural intelligence. We agree with Binz et al. that many supposedly "model-based" behaviours may be better explained by meta-learning than by classical models. We argue that this invites us to revisit our neural theories of problem solving and goal-directed planning.

The most impressive feats of natural intelligence are the most unfathomable. Caledonian crows fashion hooks to retrieve grubs, honey badgers build ladders to escape from enclosures, and humans have worked out how to split the atom (de Waal, 2016). Humans and other animals are capable of remarkable feats of problem solving in open-ended environments, but we lack computational theories of how this might be achieved (Summerfield, 2022). In the target article, Binz et al. introduce neuroscientists to an exciting new tool: Deep meta-learning. This computational approach provides an interesting candidate solution for some of nature's most startling and puzzling behaviours.

Across the twentieth century, superlative intelligence was synonymous with a capacity for planning, so early AI researchers believed that if a machine ever vanquished a human at chess, then AI would have been solved. Classical models conceive of the world as a list of states and their transition probabilities; planning requires efficient *search*, or mental exploration of possible pathways to reach a goal. Neuroscientists still lean heavily on these classical models to understand how rodents and primates solve problems like navigating to a spatial destination, assuming that these rely on "model-based" search or forms of offline rumination (Daw & Dayan, 2014). However, in contemporary machine learning, new explanations of complex sequential behaviours are emerging.

Today – nearly three decades since the first electronic chess grandmaster – AI research is dominated by deep network models, which exploit massive training datasets to learn complex functions mapping inputs onto outputs. In fact, in machine learning, explicitly model-based solutions to open-ended problems have not lived up to their promise. To give one example: In 1997, the chess program Deep Blue defeated the world champion Garry Kasparov using tree search with an alpha–beta search algorithm, ushering in an era where computers played stronger chess than people. In 2017, DeepMind's hybrid network AlphaZero defeated chess computer champion Stockfish by augmenting the search algorithm (Monte Carlo Tree Search) with a deep neural network that learned from (self-)play to evaluate board positions (Silver et al., 2018). In early 2024, performance comparable to the best human players was achieved using a deep network alone without search, thanks to computational innovation (transformer networks) and increasing scale (millions of parameters). AI research has implied that when big brains are exposed to big data, explicit forms of lookahead play a limited role in their success. The authors encapsulate this view with a quote attributed to 1920s grandmaster José Raúl Capablanca: "*I see only one move ahead, but it is always the correct one*" (Ruoss et al., 2024).

Deep meta-learning applies powerful function approximation to sequential decision problems, where optimal policies may involve forms of exploration or active hypothesis testing to meet long-term objectives. In the domain of reinforcement learning (RL), deep meta-RL can account for human behaviours on benchmark problems thought to tap into model-based inference or planning, such as the "two-step" decision task, without invoking the need for search. This is because a deep neural network equipped with a stateful activation memory, and meta-trained to a wide range of sequential decision problems, can learn a policy that is intrinsically cognitively flexible. It learns to react on the fly to the twists and turns of novel sequential environments, and thus produce the sorts of behaviours that were previously thought to be possible only with model-based forms of inference. Paradoxically, although "meta-learning" means "learning to learn," inner loop learning can occur in "frozen" networks – those without parameter updates. This offers a plausible model of how recurrent neural systems for memory and control, housed in prefrontal cortex, allow us to solve problems we have never seen before without explicit forms of search (Wang et al., 2018).

In AI research today, the most successful deep meta-learning systems are large transformer-based networks that are trained to complete sequences of tokenised natural language (Large Language Models or LLMs). These networks learn semantic and syntactic patterns that allow them to solve a very open-ended problem – constructing a relevant, coherent sentence. Researchers working with LLMs today call deep meta-learning "in context learning" because instead of using recurrent memory, transformers are purely feedforward networks that rely on autoregression – past outputs are fed back in as inputs, providing a context on which to condition the generative process. Although transformers do not resemble plausible neural algorithms, their striking success has opened up new questions concerning neural computation. For example, in-context learning proceeds faster when exemplar ordering is structured rather than random, like human learning but unlike traditional in-weight learning (Chan et al., 2022b), and "in-context learning" may be better suited to rule learning than in-weight learning (Chan et al., 2022a). Meta-learning thus offers new tools for psychologists and neuroscientists interested in biological learning and memory in natural agents.

Binz et al. argue that we should take deep meta-learning seriously as an alternative to Bayesian decision theory. We would go further, arguing that meta-learning is a candidate general theory for flexible cognition. It explains why executive function improves dramatically with experience (Ericsson & Charness, 1994). Unlike Deep Blue, human world chess no.1 Magnus Carlsen has improved since his first game at the age of five. Purely search-based accounts flexible cognition are obliged to propose that performance will plateau as soon as the transition function (e.g., the rules of chess) are fully mastered, or else posit unexplained ways in which search policies deepen or otherwise mutate with practice (Van Opheusden et al., 2023). In a world where states are heterogeneous and noisy, deep meta-learning explains how we can generalise sequential behaviours to novel states. In a world where speed and processing power are at a premium, meta-learning shifts the burden of inference to the training period, allowing for fast and efficient online computation. Meta-learning is a general theory of natural intelligence that is – more than classical counterpart – fit for the real world.

Undoubtedly, humans and other animals do engage in explicit forms of planning, especially when the stakes are high. But many sequential behaviours that are thought to index this ability may rely more on deep meta-learning than classical planning.

Financial support. This work was supported by Fondation Pour l'Audition FPA RD-2021-2 (J. P.-L.), Institute for Language, Communication, and the Brain ILCB (J. P.-L.) and Wellcome Trust Discovery Award (227928/Z/23/Z) to (C. S.).

Competing interests. None.

References

Chan, S. C. Y., Dasgupta, I., Kim, J., Kumaran, D., Lampinen, A. K., & Hill, F. (2022a). Transformers generalize differently from information stored in context vs in weights. https://doi.org/10.48550/ARXIV.2210.05675

Chan, S. C. Y., Santoro, A., Lampinen, A. K., Wang, J. X., Singh, A., Richemond, P. H., … Hill, F. (2022b). Data distributional properties drive emergent in-context learning in transformers. https://doi.org/10.48550/ARXIV.2205.05055

Daw, N. D., & Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B*, *369*, 20130478. https://doi.org/10.1098/rstb.2013.0478

de Waal, F. (2016). *Are we smart enough to know how smart animals are?* W. W. Norton & Company.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist*, *49*, 725–747. https://doi.org/10.1037/0003-066X.49.8.725

Ruoss, A., Delétang, G., Medapati, S., Grau-Moya, J., Wenliang, L. K., Catt, E., … Genewein, T. (2024). Grandmaster-level chess without search.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., … Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*, 1140–1144. https://doi.org/10.1126/science.aar6404

Summerfield, C. (2022). *Natural general intelligence: How understanding the brain can help us build A1*. Oxford University Press.

Van Opheusden, B., Kuperwajs, I., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2023). Expertise increases planning depth in human gameplay. *Nature*, *618*, 1000–1005. https://doi.org/10.1038/s41586-023-06124-2

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J.Z., … Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*, 860–868. https://doi.org/10.1038/s41593-018-0147-8

The hard problem of meta-learning is what-to-learn

Yosef Prat* and Ehud Lamm

The Cohn Institute for History and Philosophy of Science and Ideas, Tel Aviv University, Tel Aviv, Israel
yosefprat@gmail.com
ehudlamm@post.tau.ac.il
https://www.ehudlamm.com

*Corresponding author.

doi:10.1017/S0140525X24000268, e161

Abstract

Binz et al. highlight the potential of meta-learning to greatly enhance the flexibility of AI algorithms, as well as to approximate human behavior more accurately than traditional learning methods. We wish to emphasize a basic problem that lies underneath these two objectives, and in turn suggest another perspective of the required notion of "meta" in meta-learning: knowing what to learn.

We postulate that the hard problem in (natural or artificial) intelligence is the question of "what to learn?". At the fundamental level, this meta-question is resolved in nature by the evolutionary process. The question of "how to learn?", which is the focus of the meta-learning framework that is presented in the target article, is not especially easy as well, but it can be captured by devising specific training structures and relevant optimization tasks. In general, it requires to specify the "search space" of possible learning strategies. The hard problem of learning, however, is the identification of the learning task itself (i.e., what, and if, to learn) (Niv, 2019). For instance, a real-life learner observing several specimens of some unknown insect species (following the example in the target article) must first somehow realize that she is required to evaluate the average length of that species, before she begins to tune her evaluation strategies. This is indeed a different meta-task than presented by Binz et al., but its solution is mandatory for any artificial (somewhat-)general intelligence, and it is regularly handled by the brain (Roli, Jaeger, & Kauffman, 2022).

In the quest to devise domain-general learning models, Binz et al. correctly identify the need for diverse (and maybe realistic) training sets. Training a model on many different tasks can achieve high performance in all of them, and maybe even in unrelated, but similar, tasks. Yet, the model will always be constrained by the task-space spanned by its training sets. The major challenge does not lie in amplifying the dimensionality, or variability, of the learned problem, but rather in determining the appropriate objective function. Here, we may be inspired by the observation that biological brains have in general not evolved for their ability to solve a specific task, but, rather, are shaped by the overall success of the organism. On the one hand, evolutionary success obscures the objective of each specific task, since it depends on long-term benefits that are not always clearly related to short-term behavior. On the other hand, evolutionary success is a broader optimization challenge. A generic model that can both solve a maze and evaluate the average length of a newly identified insect, without being trained specifically on these tasks, must solve the hard problem of what-to-learn in a given context. To build a model that addresses this challenge we cannot handcraft the utility function (or error measurements) of each task separately. The meta-learning requirement thus becomes to learn how to identify the utility in learning, or in performing, each of the given tasks, and more broadly, to identify the task itself. Thus, it is constraining to use training sets and error functions that provide the learner with "correct" answers or feedback for each task separately, as is typically done in supervised, semi-supervised, and reinforcement machine learning. The biological brain is overall domain-general since it is not guided by a task-specific "utility function." Domain-specific processes, such as, maybe, those suggested to process language (Fedorenko & Blank, 2020), demonstrate cases in which natural selection narrowed or "optimized" the task of finding what-to-learn. Other indications may include modularity (Ellefsen, Mouret, & Clune, 2015; Sporns & Betzel, 2016), alongside sensory adaptations (Warrant, 2016), attention biases (Niv et al., 2015), and data acquisition mechanisms (Lotem & Halpern, 2012). Furthermore, in humans, cultural evolution may also adjust task specificity (Heyes, 2018).

The evolutionary process may also explain the limitations of treating cognition as rational, or optimal. Binz et al. suggest that unrealistic aspects of Bayesian models can be mitigated using resource constraints, for which the offered meta-learning framework is suitable. The problem, however, is that human (and other animal) behavior is not straightforwardly rational, and often appears to defy Bayesian optimization (Tversky & Kahneman, 1981). Moreover, this may not be due to limited resources but because the success of living creatures is determined evolutionarily, rather than by immediate outcomes (Houston, McNamara, & Steer, 2007). When behavioral objectives are considered on an evolutionary scale, it may be revealed that they are (locally) optimal (Kacelnik, 2006), and this includes behaviors that depend upon learning, as is generally assumed in behavioral ecology. When tasks for which learning is evolutionarily beneficial end up being learned (i.e., when those individuals who learn have higher fitness), natural selection resolves the meta-learning hard problem of what-to-learn (Dunlap & Stephens, 2016). This may bias the things that animals are able to learn, by shaping the parameter search-space (Prat, Bshary, & Lotem, 2022), maybe of the outer learning loop described by Binz et al. These biases are often addressed in the biological learning literature as sub-problems of the what-to-learn problem, and include when-to-learn or from whom-to-learn (Laland, 2004).

We suggest that further advancements in meta-learning thinking require addressing the hard problem of learning as one of their aims. Inspired by (human and nonhuman) biological brains, this should be done by devising overarching objectives for learning algorithms that will enable them to learn what are the learning tasks. In nature, evolution provides some of the solution. Yet, it is not necessary to mimic the evolutionary process per se, but only to acknowledge the generality of evolutionary optimization in the natural world. To this end, it may be better to aspire to simulate nonhuman-animal behavioral studies, rather than psychological assays, since nonhuman animals are trained with no description of the boundaries of their task – they need to realize it by themselves (e.g., when a sparrow learns to relate sand color to food [Ben-Oren, Truskanov, & Lotem, 2022]). Thus, these studies usually contain a direct meta-learning challenge that requires solving the problem of what-to-learn.

Acknowledgements. We thank Yoav Ram for insightful comments on a previous version of the manuscript.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

Ben-Oren, Y., Truskanov, N., & Lotem, A. (2022). House sparrows use learned information selectively based on whether reward is hidden or visible. *Animal Cognition*, *25*(6), 1545–1555. https://doi.org/10.1007/s10071-022-01637-1

Dunlap, A. S., & Stephens, D. W. (2016). Reliability, uncertainty, and costs in the evolution of animal learning. *Current Opinion in Behavioral Sciences*, *12*, 73–79. https://doi.org/https://doi.org/10.1016/j.cobeha.2016.09.010

Ellefsen, K. O., Mouret, J.-B., & Clune, J. (2015). Neural modularity helps organisms evolve to learn new skills without forgetting old skills. *PLoS Computational Biology*, *11*(4), e1004128. https://doi.org/doi.org/10.1371/journal.pcbi.1004128

Fedorenko, E., & Blank, I. A. (2020). Broca's area is not a natural kind. *Trends in Cognitive Sciences*, *24*(4), 270–284. https://doi.org/10.1016/j.tics.2020.01.001

Heyes, C. (2018). *Cognitive gadgets: The cultural evolution of thinking*. Harvard University Press. https://doi.org/10.2307/j.ctv24trbqx

Houston, A. I., McNamara, J. M., & Steer, M. D. (2007). Do we expect natural selection to produce rational behaviour? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1531–1543. https://doi.org/10.1098/rstb.2007.2051

Kacelnik, A. (2006). Meanings of rationality. In S. Hurley & M. Nudds (Eds.), *Rational animals?* (pp. 87–106). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198528272.003.0002

Laland, K. N. (2004). Social learning strategies. *Animal Learning & Behavior*, *32*(1), 4–14. https://doi.org/10.3758/BF03196002

Lotem, A., & Halpern, J. Y. (2012). Coevolution of learning and data-acquisition mechanisms: A model for cognitive evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1603), 2686–2694. https://doi.org/10.1098/rstb.2012.0213

Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, *22*(10), 1544–1553. https://doi.org/10.1038/s41593-019-0470-8

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *The Journal of Neuroscience*, *35*(21), 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015

Prat, Y., Bshary, R., & Lotem, A. (2022). Modelling how cleaner fish approach an ephemeral reward task demonstrates a role for ecologically tuned chunking in the evolution of advanced cognition. *PLoS Biology*, *20*(1), e3001519. https://doi.org/10.1371/journal.pbio.3001519

Roli, A., Jaeger, J., & Kauffman, S. A. (2022). How organisms come to know the world: Fundamental limits on artificial general intelligence. *Frontiers in Ecology and Evolution*, *9*, 806283. https://doi.org/10.3389/fevo.2021.806283

Sporns, O., & Betzel, R. F. (2016). Modular brain networks. *Annual Review of Psychology*, *67*(1), 613–640. https://doi.org/10.1146/annurev-psych-122414-033634

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458. https://doi.org/10.1126/science.7455683

Warrant, E. J. (2016). Sensory matched filters. *Current Biology*, *26*(20), R976–R980. https://doi.org/https://doi.org/10.1016/j.cub.2016.05.042

Is human compositionality meta-learned?

Jacob Russin[a,b], Sam Whitman McGrath[c], Ellie Pavlick[a] and Michael J. Frank[d]*

[a]Department of Computer Science, Brown University, Providence, RI, USA; [b]Department of Cognitive and Psychological Sciences, Brown University, Providence, RI, USA; [c]Department of Philosophy, Brown University, Providence, RI, USA and [d]Department of Cognitive and Psychological Sciences, Carney Institute for Brain Science, Brown University, Providence, RI, USA
jake_russin@brown.edu
sam_mcgrath1@brown.edu
ellie_pavlick@brown.edu
michael_frank@brown.edu
https://jlrussin.github.io/
https://scholar.google.com/citations?user=B3b7kAYAAAAJ&hl=en
https://cs.brown.edu/people/epavlick/
http://ski.clps.brown.edu/

*Corresponding author.

doi:10.1017/S0140525X24000189, e162

Abstract

Recent studies suggest that meta-learning may provide an original solution to an enduring puzzle about whether neural networks can explain compositionality – in particular, by raising the prospect that compositionality can be understood as an emergent property of an inner-loop learning algorithm. We elaborate on this hypothesis and consider its empirical predictions regarding the neural mechanisms and development of human compositionality.

Binz et al. review recent meta-learned models that can reproduce human-like compositional generalization behaviors (Lake & Baroni, 2023), but they stop short of endorsing meta-learning as a theoretical framework for understanding human compositionality. Here, we elaborate on this proposal, articulating the hypothesis that human compositionality can be understood as an emergent property of an inner-loop, in-context learning algorithm that is itself meta-learned.

Compositionality has played a key theoretical role in cognitive science since its inception (Chomsky, 1957), providing an explanation for human systematic and productive generalization behaviors. These phenomena are readily explained by the compositionality of classical cognitive architectures, as the design of their symbolic representations and structure-sensitive operations intrinsically guarantees that they can redeploy familiar constituents in novel constructions (Fodor & Pylyshyn, 1988). It has been argued that neural networks are in principle incapable of playing the same explanatory role because they lack these architectural features (Fodor & Pylyshyn, 1988; Marcus, 1998).

Much work has explored inductive biases that might encourage compositionality to emerge in neural networks (Russin, Jo, O'Reilly, & Bengio, 2020a; Smolensky, 1990; Webb et al., 2024), but meta-learning offers an original solution to the puzzle. As Binz et al. emphasize, when an inner-loop, in-context learning algorithm emerges within the activation dynamics of a meta-learning neural network, it can have fundamentally different properties than the outer-loop algorithm. Thus, even if the outer-loop algorithm lacks these inductive biases, the network may nevertheless implement an emergent in-context learning algorithm that embodies them implicitly.

Lake and Baroni (2023) have shown that such an inner-loop algorithm can pass tests of compositionality that standard neural networks fail (Lake & Baroni, 2018). The question, then, is whether such networks can serve as explanatory models of human compositional generalization. Can we think of human compositionality as an emergent property of an inner-loop, in-context learning algorithm? How might we evaluate such a hypothesis? Here, we consider two independent aspects of this proposal: First, its implications for neural mechanisms, and second, for development.

One straightforward mechanistic prediction is that employing inner-loop, in-context learning mechanisms, rather than outer-loop learning mechanisms, should facilitate compositional generalization behaviors. Cognitive and computational neuroscience provides empirical support for this prediction. Cognitive control – the ability to overcome existing prepotent responses and to flexibly adapt to arbitrary goals (Miller & Cohen, 2001) – is an important capacity for human in-context learning. The neural mechanisms known to be involved in cognitive control, such as working memory, gating, and top-down modulation in the prefrontal cortex (Miller & Cohen, 2001; O'Reilly & Frank, 2006; Russin, O'Reilly, & Bengio, 2020b), are also thought to be essential to compositional abilities such as inferring and applying rules (Calderon, Verguts, & Frank, 2022; Collins & Frank, 2013; Frank & Badre, 2012; Kriete, Noelle, Cohen, & O'Reilly, 2013), deductive and inductive reasoning (Crescentini et al., 2011; Goel, 2007), and processing complex syntax (Thompson-Schill, 2005). Thus, a shared set of neural mechanisms may underlie both in-context learning and compositionality in humans, lending support to the meta-learning hypothesis.

A second, independent prediction is a developmental one – that human compositional generalization abilities are themselves meta-learned over the course of development. Adults come into any psychological experiment equipped with a wealth of prior experience. The meta-learning hypothesis predicts that this includes experiences encouraging the adoption of more compositional learning strategies (i.e., ones sensitive to implicit compositional structure). In general, children exhibit a developmental trajectory consistent with this hypothesis. Older children learn new tasks more efficiently (Bergelson, 2020), especially when these tasks involve cognitive capacities essential to in-context learning, such as working memory and executive functions (Munakata, Snyder, & Chatham, 2012). Furthermore, children improve throughout development on tasks involving the composition of rules (Piantadosi & Aslin, 2016; Piantadosi, Palmeri, & Aslin, 2018).

Innate mechanisms or inductive biases may still be required to successfully meta-learn a compositional inner-loop algorithm in the first place. Indeed, studies in machine learning have shown that architecture seems to be an important factor in determining whether in-context learning capabilities emerge (Chan et al., 2022). Similarly, findings from cognitive and computational neuroscience have emphasized the importance of architectural features such as prefrontal gating mechanisms for the emergence of abstract representations that could mediate subsequent in-context generalization abilities (Collins & Frank, 2013; Frank & Badre, 2012; Kriete et al., 2013; Rougier, Noelle, Braver, Cohen, & O'Reilly, 2005). These inductive biases can also explain incidental hierarchical rule learning and generalization in infants (Werchan, Collins, Frank, & Amso, 2015, 2016). Thus, a combination of innate architectural features and meta-learning experiences may be necessary for human compositionality to emerge.

The meta-learning datasets used in previous modeling efforts have typically been developmentally unrealistic because they have been contrived to engender narrow compositional generalization abilities that are specific to a particular type of task. Could meta-learning in less explicitly structured learning scenarios lead to the acquisition of broader compositional generalization abilities? This question deserves careful empirical study, but we may draw a preliminary insight from the success of large language models (Brown et al., 2020), which develop in-context learning abilities (von Oswald et al., 2023; Xie, Raghunathan, Liang, & Ma, 2022) that in some cases exhibit human-like compositionality (Webb, Holyoak, & Lu, 2022; Wei et al., 2023; Zhou et al., 2022). Unlike models explicitly designed for meta-learning, large language models are trained to predict the next token on very large datasets of unstructured text. These datasets contain more language data than humans are exposed to in an entire lifetime (Linzen & Baroni, 2021), so future work needs to investigate what kinds of inductive biases are necessary to improve their sample efficiency. However, these models provide proof of concept that neural networks can develop compositional in-context learning algorithms by training on relatively unstructured data.

Binz et al. shy away from a robust commitment to meta-learning as a theoretical framework, instead emphasizing its utility as a methodological tool. Here, we have demonstrated how the meta-learning perspective on human compositionality can generate testable empirical hypotheses about underlying mechanisms and developmental trajectory. If such a research program bears fruit, it will elevate meta-learning from a useful tool to a novel cognitive theory.

Financial support. M. J. F. is supported by ONR grant N00014-23-1-2792. E. P. and J. R. are supported by COBRE grant no. 5P20GM103645-10.

Competing interest. None.

References

Bergelson, E. (2020). The comprehension boost in early word learning: Older infants are better learners. *Child Development Perspectives*, *14*(3), 142–149. https://doi.org/10.1111/cdep.12373

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P.. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Calderon, C. B., Verguts, T., & Frank, M. J. (2022). Thunderstruck: The ACDC model of flexible sequences and rhythms in recurrent neural circuits. *PLoS Computational Biology*, *18*(2), e1009854. https://doi.org/10.1371/journal.pcbi.1009854

Chan, S. C. Y., Santoro, A., Lampinen, A. K., Wang, J. X., Singh, A., Richemond, P. H., … Hill, F. (2022). Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, *35*, 18878–18891. https://papers.nips.cc/paper_files/paper/2022/hash/77c6ccacfd9962e2307fc64680fc5ace-Abstract-Conference.html

Chomsky, N. (Ed.). (1957). *Syntactic structures*. Mouton & Co.

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering and generalizing task-set structure. *Psychological Review*, *120*(1), 190–229. https://doi.org/10.1037/a0030852

Crescentini, C., Seyed-Allaei, S., De Pisapia, N., Jovicich, J., Amati, D., & Shallice, T. (2011). Mechanisms of rule acquisition and rule following in inductive reasoning. *Journal of Neuroscience*, *31*(21), 7763–7774. https://doi.org/10.1523/JNEUROSCI.4579-10.2011

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1–2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Frank, M. J., & Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: Computational analysis. *Cerebral Cortex*, *22*(3), 509–526. https://doi.org/10.1093/cercor/bhr114

Goel, V. (2007). Anatomy of deductive reasoning. *Trends in Cognitive Sciences*, *11*(10), 435–441. https://doi.org/10.1016/j.tics.2007.09.003

Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(41), 16390–16395. https://doi.org/10.1073/pnas.1303547110

Lake, B. M., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In J. G. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 2879–2888). PMLR. http://proceedings.mlr.press/v80/lake18a.html

Lake, B. M., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, *623*, 1–7. https://doi.org/10.1038/s41586-023-06668-3

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*(1), 195–212. https://doi.org/10.1146/annurev-linguistics-032020-051035

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, *37*(3), 243–282. https://doi.org/10.1006/cogp.1998.0694

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Munakata, Y., Snyder, H. R., & Chatham, C. H. (2012). Developing cognitive control: Three key transitions. *Current Directions in Psychological Science*, *21*(2), 71–77. https://doi.org/10.1177/0963721412436807

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, *18* (2), 283–328. https://doi.org/10.1162/089976606775093909

Piantadosi, S., & Aslin, R. (2016). Compositional reasoning in early childhood. *PLoS ONE*, *11*(9), e0147734. https://doi.org/10.1371/journal.pone.0147734

Piantadosi, S. T., Palmeri, H., & Aslin, R. (2018). Limits on composition of conceptual operations in 9-month-olds. *Infancy*, *23*(3), 310–324. https://doi.org/10.1111/infa.12225

Rougier, N. P., Noelle, D., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and the flexibility of cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(20), 7338–7343.

Russin, J., Jo, J., O'Reilly, R. C., & Bengio, Y. (2020a). Systematicity in a recurrent neural network by factorizing syntax and semantics. *Proceedings for the 42nd Annual Meeting of the Cognitive Science Society*, 7. https://cognitivesciencesociety.org/cogsci20/papers/0027/0027.pdf

Russin, J., O'Reilly, R. C., & Bengio, Y. (2020b). Deep learning needs a prefrontal cortex. In Bridging AI and Cognitive Science (BAICS) Workshop, ICLR, 2020, 11.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1–2), 159–216. https://doi.org/10.1016/0004-3702(90)90007-M

Thompson-Schill, S. L. (2005). Dissecting the language organ: A new look at the role of Broca's area in language processing. In Anne Cutler (Ed.), *Twenty-first century psycholinguistics* (1st ed., Vol. 1, pp. 1–18). Routledge.

von Oswald, J., Niklasson, E., Schlegel, M., Kobayashi, S., Zucchet, N., Scherrer, N., … Sacramento, J. (2023). Uncovering mesa-optimization algorithms in transformers (arXiv:2309.05858). arXiv. https://doi.org/10.48550/arXiv.2309.05858

Webb, T., Frankland, S. M., Altabaa, A., Krishnamurthy, K., Campbell, D., Russin, J., … Cohen, J. D. (2024). The relational bottleneck as an inductive bias for efficient abstraction (arXiv:2309.06629). arXiv. http://arxiv.org/abs/2309.06629

Webb, T., Holyoak, K. J., & Lu, H. (2022). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9). https://doi.org/10.1038/s41562-023-01659-w

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., … Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837. https://papers.nips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

Werchan, D. M., Collins, A. G. E., Frank, M. J., & Amso, D. (2015). 8-Month-old infants spontaneously learn and generalize hierarchical rules. *Psychological Science*, *26*(6), 805–815. https://doi.org/10.1177/0956797615571442

Werchan, D. M., Collins, A. G. E., Frank, M. J., & Amso, D. (2016). Role of prefrontal cortex in learning and generalizing hierarchical rules in 8-month-old infants. *The Journal of Neuroscience*, *36*(40), 10314–10322. https://doi.org/10.1523/JNEUROSCI.1351-16.2016

Xie, S. M., Raghunathan, A., Liang, P., & Ma, T. (2022). An explanation of in-context learning as implicit Bayesian inference. *International Conference on Learning Representations*. https://openreview.net/pdf?id=RdJVFCHjUMI

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., … Chi, E. (2022). Least-to-most prompting enables complex reasoning in large language models. *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=WZH7099tgfM

Combining meta-learned models with process models of cognition

Adam N. Sanborn*, Haijiang Yan and Christian Tsvetkov

Department of Psychology, University of Warwick, Coventry, UK
a.n.sanborn@warwick.ac.uk
haijiang.yan@warwick.ac.uk
chris.tsvetkov@warwick.ac.uk
https://go.warwick.ac.uk/adamsanborn

*Corresponding author.

doi:10.1017/S0140525X24000165, e163

Abstract

Meta-learned models of cognition make optimal predictions for the actual stimuli presented to participants, but investigating judgment biases by constraining neural networks will be unwieldy. We suggest combining them with cognitive process models, which are more intuitive and explain biases. Rational process models, those that can sequentially sample from the posterior distributions produced by meta-learned models, seem a natural fit.

Meta-learned models of cognition offer an exciting opportunity to address a central weakness of current cognitive models, whether Bayesian or not: Cognitive models generally do not "see" the experimental stimuli shown to participants. Experimenters instead feed models low-dimensional descriptions of the stimuli, which are often in terms of the psychological features imagined by the experimenter, or sometimes are the psychological descriptions that best fit participants' judgments (e.g., stimulus similarity judgments; Nosofsky, Sanders, Meagher, & Douglas, 2018).

For example, in studies of probability judgment, participants have been asked to judge the probability that "Bill plays jazz for a hobby" after having been given the description, "Bill is 34 years old. He is intelligent, but unimaginative, compulsive, and generally lifeless. In school, he was strong in mathematics but weak in social studies and humanities" (Tversky & Kahneman, 1983). Current probability judgment models reduce these descriptions down to a single unknown number, and attempt to find the latent probability that best fits the data (e.g., Zhu, Sanborn, & Chater, 2020).

Models trained on the underlying statistics of the environment, as meta-learned models are, can bypass this need to infer a latent variable, instead making predictions from the actual descriptions used. Indeed, even relatively simple models of semantics that locate phrases in a vector space produce judgments that correlate with the probabilities experimental participants give (Bhatia, 2017). Meta-learned models could thus explain a great deal of the variability in human behavior, and allow experimenters to generalize beyond the stimuli shown to participants.

However, used as descriptive models, normative meta-learned models of cognition inherit a fundamental problem from the Bayesian approach: People's reliable deviations from normative behavior. One compelling line of research shows that probability judgments are incoherent in a way that Bayesian models are not. Using the above example of Bill, Tversky and Kahneman (1983) found participants ranked the probability of "Bill is an accountant who plays jazz for a hobby" as higher than that of "Bill plays jazz for a hobby." This violates the extension rule of probability because the set of all accountants who play jazz for a hobby is a subset of all people who play jazz for a hobby, no matter how Bill is described.

The target article discusses constraining meta-learned models to better describe behavior, such as reducing the number of hidden units or restricting the representational fidelity of units. These manipulations have produced a surprising and interesting range of biases, including stochastic and incoherent probability judgments (Dasgupta, Schulz, Tenenbaum, & Gershman, 2020). However, this is just the start to explaining human biases. Even a single bias such as the conjunction fallacy has intricacies, such as the higher rate of conjunction fallacies when choosing versus estimating (Wedell & Moro, 2008), and greater variability in judgments of conjunctions than those of simple events (Costello & Watts, 2017).

Cognitive process models aim to explain these biases in detail. For conjunction fallacies, a variety of well-supported models exist, based on ideas such as participants sampling events with noise in the retrieval process (Costello & Watts, 2014), or by sacrificing probabilistic coherence to improve judgment accuracy based on samples (Zhu et al., 2020), or by representing conjunctions as a weighted average of simple events (Juslin, Nilsson, & Winman, 2009), or by using quantum probability (Busemeyer, Pothos, Franco, & Trueblood, 2011). These kinds of models capture many details of the empirical effects, through simple and intuitive mechanisms like adjusting the amount of noise or number of samples, which helps identify experiments to distinguish between them.

Mechanistically modifying meta-learned models to explain cognitive biases to the level cognitive process models do appears difficult. While changes to network structure are powerful ways to induce different biases that could identify implementation-level constraints in the brain, the effects of these kinds of changes are generally hard to intuit, while training constrained meta-learning models to test different manipulations will be slow and computationally expensive. Thus, it will be challenging to reproduce existing biases in detail or to design effective experiments for testing these constraints.

Combining meta-learned models with cognitive process models is more promising. One possibility is to have meta-learned models act as a "front end" that takes stimuli and converts them to a feature-based representation, which is then operated on by a cognitive process model. The parameters of the cognitive process model could be fit to human data, or potentially the cognitive process model could be encoded into the network (e.g., Peterson, Bourgin, Agrawal, Reichman, & Griffiths, 2021), and meta-learning could be done on the front end and the cognitive process parameters end-to-end.

However, as meta-learned models of cognition produce posterior predictive distributions, rational process models offer a straightforward connection that does not require retraining meta-learned models. Rational process models do not directly use a posterior predictive distribution, but instead assume that the posterior predictive distribution is approximated (i.e., using the posterior mean, posterior median, or other summary statistic depending on task), most often using a statistical sampling algorithm (Griffiths, Vul, & Sanborn, 2012). Such a model can explain details of the conjunction fallacy, and also a wide range of other biases, such as stochastic choice, anchoring and repulsion effects in estimates, long-range autocorrelations in judgment, and the flaws in random sequence generation (Castillo, León-Villagrá, Chater, & Sanborn, 2024; Spicer, Zhu, Chater, & Sanborn, 2022; Vul, Goodman, Griffiths, & Tenenbaum, 2014; Zhu, León-Villagrá, Chater, & Sanborn, 2022, 2023). What these models have lacked, however, is a principled way to construct the posterior predictive distribution from environmental statistics, and here meta-learned models offer that exciting possibility.

While rational process models offer what we think is a natural choice for integration, any sort of combination with existing cognitive models offers benefits. Being able to explain both the details of biases as cognitive process models do, as well as showing sensitivity to actual stimuli is a powerful combination that moves toward the long-standing goal of a general model of cognition. Overall we see meta-learned models of cognition as not supplanting existing cognitive models, but as a way to make them much more powerful and relevant to understanding and predicting behavior.

Acknowledgments. None.

Financial support. A. N. S. and C. T. were supported by a European Research Council consolidator grant (817492 – SAMPLING). H. Y. was supported by a Chancellor's International Scholarship from the University of Warwick.

Competing interest. None.

References

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1–20. http://dx.doi.org/10.1037/rev0000047

Busemeyer, J. R., Pothos, E. M., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological Review*, *118*(2), 193–218. https://doi.org/10.1037/a0022542

Castillo, L., León-Villagrá, P., Chater, N., & Sanborn, A. (2024). Explaining the flaws in human random generation as local sampling with momentum. *PLoS Computational Biology*, *20*(1), e1011739. https://doi.org/10.1371/journal.pcbi.1011739

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463–480. https://doi.org/10.1037/a0037010

Costello, F., & Watts, P. (2017). Explaining high conjunction fallacy rates: The probability theory plus noise account. *Journal of Behavioral Decision Making*, *30*(2), 304–321. https://dx.doi.org/10.1002/bdm.1936

Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review*, *127*(3), 412–441. https://doi.org/10.1037/rev0000178

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263–268. https://doi.org/10.1177/0963721412447619

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review*, *116*(4), 856–874. https://doi.org/10.1037/a0016979

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2018). Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods*, *50*, 530–556. https://doi.org/10.3758/s13428-017-0884-8

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214. https://doi.org/10.1126/science.abe2629

Spicer, J., Zhu, J. Q., Chater, N., & Sanborn, A. N. (2022). Perceptual and cognitive judgments show both anchoring and repulsion. *Psychological Science*, *33*(9), 1395–1407. https://doi.org/10.1177/09567976221089599

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637. https://doi.org/10.1111/cogs.12101

Wedell, D. H., & Moro, R. (2008). Testing boundary conditions for the conjunction fallacy: Effects of response mode, conceptual focus, and problem type. *Cognition*, *107*(1), 105–136. https://doi.org/10.1016/j.cognition.2007.08.003

Zhu, J. Q., León-Villagrá, P., Chater, N., & Sanborn, A. N. (2022). Understanding the structure of cognitive noise. *PLoS Computational Biology*, *18*(8), e1010312. https://doi.org/10.1371/journal.pcbi.1010312

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The Bayesian sampler: Generic Bayesian inference causes incoherence in human probability judgments. *Psychological Review*, *127*(5), 719–748. https://doi.org/10.1037/rev0000190

Zhu, J.-Q., Sundh, J., Spicer, J., Chater, N., & Sanborn, A. N. (2023). The autocorrelated Bayesian sampler: A rational process for probability judgments, estimates, confidence intervals, choices, confidence judgments, and response times. *Psychological Review*, *131*(2), 456–493. https://doi.org/10.1037/rev0000427

Linking meta-learning to meta-structure

Malte Schilling[a]*, Helge J. Ritter[b] and Frank W. Ohl[c,d]

[a]Autonomous Intelligent Systems Group, Computer Science Department, University of Münster, Münster, Germany; [b]Neuroinformatics Group Faculty of Technology/CITEC, Bielefeld University, Bielefeld, Germany; [c]Department of Systems Physiology of Learning, Leibniz Institute for Neurobiology, Magdeburg, Germany and [d]Institute of Biology, Otto-von-Guericke University, Magdeburg, Germany
malte.schilling@uni-muenster.de
helge@techfak.uni-bielefeld.de
frank.ohl@lin-magdeburg.de
https://www.uni-muenster.de/AISystems/
https://ni.www.techfak.uni-bielefeld.de/people/helge/
https://www.ovgu.de/Ohl.html

*Corresponding author.

doi:10.1017/S0140525X24000232, e164

Abstract

We propose that a principled understanding of meta-learning, as aimed for by the authors, benefits from linking the focus on learning with an equally strong focus on structure, which means to address the question: What are the meta-structures that can guide meta-learning?

The authors discuss meta-learning as a flexible and computationally efficient tool to generate cognitive models from training data and thereby to avoid the need for handcrafting cognitive biases as usually done in current cognitive architectures or Bayesian learning. They provide four supporting arguments as a motivation for a systematic research program on meta-learning, which they diagnose as so far largely missing. While we agree with this stance, we propose that a deeper understanding of meta-learning would benefit from complementing the focus on learning with *an equally* *strong focus on structure*, that is, to address the question: *What are the meta-structures that are decisive to shape meta-learning?*

The reasoning for our proposal derives from the authors' "Argument 3", where they argue that meta-learning makes it easy to manipulate a learning algorithm's complexity to construct resource-rational models of learning. By admitting complexity as an important control for model formation, the authors introduce structural discriminations between meta-learners. But as a scalar measure, complexity cannot avoid "collapsing" qualitatively different structures whenever these are assigned the same complexity. Therefore, we suggest extending the research program beyond scalar orderings as complexity measures: Viewing **meta-structures** as patterns of higher-order structure that are qualitatively different from each other and that offer structural-functional "modules" that can be constructed as entities in their own right and be flexibly used by a meta-learning system. This view draws close inspiration from the advocated neuroscience perspective (their "Argument 4") how constraints of the neurobiological substrate determine the emergence of specific control structures. **Meta-structures** are thus abstractable structural principles guiding the development of "substrate-level" structures in a meta-learning system. While meta-learning summarizes many learning trajectories into an overarching base learner (that can quickly specialize), meta-structure summarizes many learning priors into an overarching "base prior" (that then guides meta-learning efficiently).

Examples for such guiding meta-structures can be found in biological neural network models. As a first meta-structure, we consider **hierarchical organization**: The decomposition of actions into sub-actions on different levels of a hierarchy enables flexible recombination into different behaviors. Hierarchical organization is an established principle of biological motor control that has been applied successfully to Deep Reinforcement Learning (DRL) (Merel, Botvinick, & Wayne, 2019; Neftci & Averbeck, 2019). As a benefit, hierarchical organization enables a form of higher-level learning in which the learner can recombine modular policies into new behaviors without the need to always learn all details from scratch.

As a second example of a meta-structure: **Decentralization** serves parallelization of modules' actions, decoupling of subtasks and factorization of state spaces. Decentralization is well-investigated in motor control in animals, for example, in low-level reflexes, but it is also widely acknowledged that decentralized oscillation generating neuronal circuits are essential for locomotion (cf. Dickinson et al., 2000). While decentralization often is merely characterized as a strategy to cope with slow sensory processing, we emphasize how decentralization facilitates meta-learning. In a study on learning of motor control featuring decentralized modules for a four-legged walker, we showed how decentralization positively affected reinforcement learning on two levels (Schilling, Melnik, Ohl, Ritter, & Hammer, 2021). First, on the basic learning level, decentralization remedies the problem of exponential increase of required training runs in traditional DRL systems as the action space becomes more complex. Decentralization restricts the action space to the much lower number of local actuators, thereby reducing the dimensionality. Without the need for coordination of all control signals by a single centralized controller, the decentralized network learned stable behaviors much faster. Second, on the level of meta-learning, the trained decentralized controller appeared to learn a different, more robust, mapping when compared to a standard centralized controller: The decentralized control structure had learned to transfer previously learned aspects of motor control to entirely new terrains without the need for further context-specific training. Thus, with respect to meta-learning this structural prior (meta-structure) of decentralization proved beneficial for extrapolation of behavior and appears to learn better suited mappings for a broader range of tasks.

A further example of meta-structures can be identified in reversal learning. In **reversal learning**, an agent initially learns a mapping, for example, between certain situations and corresponding appropriate responses, and then finds itself in a situation that *requires a different stimulus response mapping* to achieve behavioral goals. While standard DRL agents learn new mappings at a reversal point from scratch, biological organisms typically solve reversal problems more effectively (Happel et al., 2014): They create already during the initial learning phase hierarchically organized representation structures that can be efficiently used for the new required mapping without learning it from scratch (Jarvers et al., 2016).

The examples imply that a common mechanism by which meta-structures support meta-learning is that of enabling a learning agent to build context. Context-building is naturally introduced by meta-learning itself, for example, as conceived in the target article: An inner-loop learner operates at the fast time scale, with temporally short-ranged contexts, while an "outer-loop process," which could be implemented either as dedicated network modules or as processes made possible on decentralized structures, works at time scales and higher levels of abstraction that allow tuning of inner learner's adaptivity (Schilling, Hammer, Ohl, Ritter, & Wiskott, 2023). This meta-structure can be imagined as recursively extendable, leading to an "onion-like" architecture providing a principled stratification of an overall learning process into a layered hierarchy of learners operating at different levels of granularity (or abstraction), with correspondingly scaled scopes of context. Such concepts can contribute to our understanding of sophisticated learning capabilities as needed by embodied agents that continuously adapt interactions of their body with the environment involving contexts at different temporal and spatial scales.

The discussion of meta-structures illustrates that meta-learning is based on abstractable structural principles that support the generation of a **compositional semantics** linking the functions of modules that emerge from learning in a given context. Meta-structures provide structural preconditions for the establishment of almost arbitrary high-level compositional semantics that can be flexibly reused and serve as building blocks of higher-level abstractions for novel problem solutions without requiring specific training in novel contexts. Together, these aspects underscore the significance of meta-structure for a research program on meta-learning.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. None.

References

Dickinson, M. H., Farley, C. T., Full, R. J., Koehl, M., Kram, R., & Lehman, S. (2000). How animals move: An integrative view. *Science 288*(5463), 100–106. https://doi.org/10.1126/science.288.5463.100

Happel, M. F., Niekisch, H., Castiblanco Rivera, L. L., Ohl, F. W., Deliano, M., & Frischknecht, R. (2014). Enhanced cognitive flexibility in reversal learning induced by removal of the extracellular matrix in auditory cortex. *Proceedings of the National Academy of Sciences*, *111*(7), 2800–2805.

Jarvers, C., Brosch, T., Brechmann, A., Woldeit, M. L., Schulz, A. L., Ohl, F. W., … Neumann, H. (2016). Reversal learning in humans and gerbils: Dynamic control network facilitates learning. *Frontiers in Neuroscience*, *10*, 535.

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

Merel, J., Botvinick, M., & Wayne, G. (2019). Hierarchical motor control in mammals and machines. *Nature Communications*, *10*(1), 5489.
Neftci, E. O., & Averbeck, B. B. (2019). Reinforcement learning in artificial and biological systems. *Nature Machine Intelligence*, *1*(3), 133–143.
Schilling, M., Hammer, B., Ohl, F. W., Ritter, H. J., & Wiskott, L. (2023). Modularity in nervous systems – a key to efficient adaptivity for deep reinforcement learning. *Cognitive Computation*, 1–16.
Schilling, M., Melnik, A., Ohl, F. W., Ritter, H. J., & Hammer, B. (2021). Decentralized control and local information for robust and adaptive decentralized deep reinforcement learning. *Neural Networks*, *144*, 699–725.

The added value of affective processes for models of human cognition and learning

Yoann Stussi[a,b]*, Daniel Dukes[a] and David Sander[a,b]

[a]Swiss Center for Affective Sciences, Campus Biotech, University of Geneva, Geneva, Switzerland and [b]Department of Psychology, FPSE, University of Geneva, Geneva, Switzerland
yoann.stussi@unige.ch
daniel.dukes@unige.ch
david.sander@unige.ch
https://www.unige.ch/fapse/e3lab/members1/senior-researchers-and-lecturers/dr-yoann-stussi
https://dukes.space/
https://www.unige.ch/fapse/e3lab/director/

*Corresponding author.

doi:10.1017/S0140525X24000207, e165

Abstract

Building on the affectivism approach, we expand on Binz et al.'s meta-learning research program by highlighting that emotion and other affective phenomena should be key to the modeling of human learning. We illustrate the added value of affective processes for models of learning across multiple domains with a focus on reinforcement learning, knowledge acquisition, and social learning.

Binz et al. bid to establish an ambitious research program to model human cognition using a meta-learning framework. They effectively illustrate the potential and advantages of meta-learned models, showcasing the ability of such models to acquire inductive biases through experience. Notably, the authors outline a compelling blueprint for how these models could foster the development of a domain-general model of human learning. Here, we seek to complement this blueprint by highlighting a key element that is left mostly unexplored in the target article: Affective processes.

Affective processes – typically emotions, feelings, motivations, moods, or attitudes – are not only inherently linked to well-being, but also drive behavioral and cognitive processes such as attention, learning, memory, and decision-making (e.g., LaBar & Cabeza, 2006; Lerner, Li, Valdesolo, & Kassam, 2015; Phelps, 2006; Pool, Brosch, Delplanque, & Sander, 2016). This grants affective processes high explanatory power in understanding human behavior and cognition, a central argument of the affectivism approach (Dukes et al., 2021). As such, we suggest that considering affective processes is pivotal to the modeling of human cognition, and especially of learning. Affective processes are indeed central to – and exert a pervasive influence on – how humans learn (e.g., Öhman & Mineka, 2001; Vollberg & Sander, 2024; Wuensch, Pool, & Sander, 2021). Below, we illustrate how emotion and other affective phenomena are central to human learning across various domains, with a particular focus on reinforcement learning, knowledge acquisition, and social learning.

Affective processes emerge as important factors in reinforcement learning. This fundamental learning process enables individuals to attribute value to states or stimuli and actions via teaching signals such as rewards and punishments. These reinforcers and their associated stimuli typically evoke affective responses, which are core components of reward-seeking and threat-related behaviors (Levy & Schiller, 2021; Stussi & Pool, 2022). Affective processes also modulate how individuals learn from reinforcers. Studies on Pavlovian conditioning – a basic form of reinforcement learning – have shown that stimuli with heightened affective relevance, such as both threat-relevant (e.g., angry faces) and positive emotional (e.g., baby faces) stimuli, are more rapidly and persistently associated with an aversive outcome than neutral stimuli (Stussi, Pourtois, & Sander, 2018; Stussi, Pourtois, Olsson, & Sander, 2021). At the computational level, these studies indicate that affective relevance modulates how individuals learn from prediction errors: Affectively relevant stimuli were associated with a lower learning rate for negative prediction errors (i.e., when the aversive outcome was expected but omitted), enhancing the persistence of their association with the aversive outcome (Stussi et al., 2018, 2021). Similarly, substantial evidence has demonstrated that individuals learn differently about positive and negative outcomes in the instrumental domain (Dorfman, Bhui, Hughes, & Gershman, 2019; Lefebvre, Lebreton, Meyniel, Bourgeois-Gironde, & Palminteri, 2017). Positively valenced prediction errors are generally associated with a higher learning rate than negatively valenced prediction errors, providing a computational correlate of such learning asymmetry (Palminteri & Lebreton, 2022). Altogether, these findings highlight that affective mechanisms shape basic reinforcement learning processes.

Affective processes likewise support epistemic learning. Both positive and negative emotions have long been studied for their roles in the encoding, consolidation, and recall of episodic memories (Levine & Pizarro, 2004), as well as in academic learning in educational settings (see Pekrun & Linnenbrink-Garcia, 2014). Epistemic emotions are the key family of emotions supporting knowledge exploration and acquisition (see Muis, Chevrier, & Singh, 2018). Emotions such as interest, curiosity, confusion, surprise, wonder, or awe are drivers of learning (e.g., Chevrier, Muis, Trevors, Pekrun, & Sinatra, 2019; Vogl, Pekrun, Murayama, & Loderer, 2020). As an illustration of the central role of epistemic emotions in learning, the "trivia questions" paradigm is typically used to understand how epistemic curiosity enhances memory. Using this paradigm, research has shown that the more participants are curious to know the response to a question (e.g., "who is the most cited psychologist of the 21st century?"), the more they later remember the response (e.g., Kang et al., 2009; Marvin & Shohamy, 2016). The impact of curiosity on knowledge exploration and acquisition, partly relying on reward-related processes, is therefore a salient example of how the intrinsic value of information can enhance learning (Murayama, 2022).

While many things can be learned by exploring one's own environment, this individual approach has its limitation. Some information simply cannot be gleaned in this way and requires input from other (human) sources (Harris & Koenig, 2006). Critically, such social learning fundamentally relies on affective processes. Social learning has historically been seen as either a non-human primate phenomenon that explains how behavior can be learned from conspecifics (Zentall & Galef, 1988), or a human cognitive developmental phenomenon concerned with learning from others' testimony (Harris, 2012). However, affective social learning not only points out that these two branches of social learning originate from the same tree (Gruber, Bazhydai, Sievers, Clément, & Dukes, 2022), but also that it is possible to learn from others' affective attitudes about the value of objects (e.g., ideas, people, customs). An important part of who we are – our values, ethics, and morality – is based on our perception, attention, and memory of interaction with and learning from others, whether or not this information is communicated ostensively (Dukes & Clément, 2017; Egyed, Király, & Gergely, 2013). And indeed, what we perceive, attend to, and remember is largely defined by how important, valuable, and affective those objects of perception, attention, and memory are. Not only do we remember what is affectively relevant to us as individuals, but others, serving as proxy relevant detectors, can also signal what is more or less relevant, to be learned or forgotten (Dukes & Clément, 2019; see also Sorce, Emde, Campos, & Klinnert, 1985).

In conclusion, affective processes play a fundamental role in learning across various domains and their consideration is key to the modeling of human learning. Given that emotions are not immutable and static but flexibly arise from the interaction between an individual and their environment (Scherer & Moors, 2019), it could be particularly enlightening to conceptualize emotion as a kind of inductive bias attuned by experience within the meta-learning framework proposed by Binz et al. Such conceptualization could offer a promising way of modeling the effects of emotion on learning, thereby providing added value to the meta-learning research agenda.

Acknowledgements. The authors thank Dr. Eva R. Pool, Professor Dr. Maël Lebreton, and Professor Dr. Fabrice Clément for insightful discussions.

Financial support. Y. S. is supported by an ERC Starting Grant (INFORL-948671) awarded to Professor Dr. Maël Lebreton.

Competing interest. None.

References

Chevrier, M., Muis, K. R., Trevors, G. J., Pekrun, R., & Sinatra, G. M. (2019). Exploring the antecedents and consequences of epistemic emotions. *Learning and Instruction*, *63*, Article 101209. https://doi.org/10.1016/j.learninstruc.2019.05.006

Dorfman, H. M., Bhui, R., Hughes, B. L., & Gershman, S. J. (2019). Causal inference about good and bad outcomes. *Psychological Science*, *30*(4), 516–525. https://doi.org/10.1177/0956797619828724

Dukes, D., Abrams, K., Adolphs, R., Ahmed, M. E., Beatty, A., Berridge, K. C., … … Sander, D. (2021). The rise of affectivism. *Nature Human Behaviour*, *5*, 816–820. https://doi.org/10.1038/s41562-021-01130-8

Dukes, D., & Clément, F. (2017). Author reply: Clarifying the importance of ostensive communication in long-life, affective social learning. *Emotion Review*, *9*(3), 267–269. https://doi.org/10.1177/1754073916679006

Dukes, D., & Clément, F. (Eds.). (2019). *Foundations of affective social learning: Conceptualizing the social transmission of value*. Cambridge University Press. https://doi.org/10.1017/9781108661362

Egyed, K., Király, I., & Gergely, G. (2013). Communicating shared knowledge in infancy. *Psychological Science*, *24*(7), 1348–1353. https://doi.org/10.1177/0956797612471952

Gruber, T., Bazhydai, M., Sievers, C., Clément, F., & Dukes, D. (2022). The ABC of social learning: Affect, behavior, and cognition. *Psychological Review*, *129*(6), 1296–1318. https://doi.org/10.1037/rev0000311

Harris, P. L. (2012). *Trusting what you're told: How children learn from others*. Harvard University Press. https://doi.org/10.4159/harvard.9780674065192

Harris, P. L., & Koenig, M. A. (2006). Trust in testimony: How children learn about science and religion. *Child Development*, *77*(3), 505–524. https://doi.org/10.1111/j.1467-8624.2006.00886.x

Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T., & Camerer, C. F. (2009). The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhanced memory. *Psychological Science*, *20*(8), 963–973. https://doi.org/10.1111/j.1467-9280.2009.02402.x

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience*, *7*, 54–64. https://doi.org/10.1038/nrn1825

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, *1*, Article 0067. https://doi.org/10.1038/s41562-017-0067

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology*, *66*, 799–823. https://doi.org/10.1146/annurev-psych-010213-115043

Levine, L. J., & Pizarro, D. A. (2004). Emotion and memory research: A grumpy overview. *Social Cognition*, *22*(5), 530–554. https://doi.org/10.1521/soco.22.5.530.50767

Levy, I., & Schiller, D. (2021). Neural computations of threat. *Trends in Cognitive Sciences*, *25*(2), 151–171. https://doi.org/10.1016/j.tics.2020.11.007

Marvin, C. B., & Shohamy, D. (2016). Curiosity and reward: Valence predicts choice and information prediction errors enhance learning. *Journal of Experimental Psychology: General*, *145*(3), 266–272. https://doi.org/10.1037/xge0000140

Muis, K. R., Chevrier, M., & Singh, C. A. (2018). The role of epistemic emotions in personal epistemology and self-regulated learning. *Educational Psychologist*, *53*(3), 165–184. https://doi.org/10.1080/00461520.2017.1421465

Murayama, K. (2022). A reward-learning framework of knowledge acquisition: An integrated account of curiosity, interest, and intrinsic-extrinsic rewards. *Psychological Review*, *129*(1), 175–198. https://doi.org/10.1037/rev0000349

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, *108*(3), 483–522. https://doi.org/10.1037/0033-295X.108.3.483

Palminteri, S., & Lebreton, M. (2022). The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in Cognitive Sciences*, *26*(7), 607–621. https://doi.org/10.1016/j.tics.2022.04.005

Pekrun, R., & Linnenbrink-Garcia, L. (Eds.). (2014). *International handbook of emotion in education*. Routledge.

Phelps, E. A. (2006). Emotion and cognition: Insights from studies of the human amygdala. *Annual Review of Psychology*, *57*, 27–53. https://doi.org/10.1146/annurev.psych.56.091103.070234

Pool, E., Brosch, T., Delplanque, S., & Sander, D. (2016). Attentional bias for positive emotional stimuli: A meta-analytic investigation. *Psychological Bulletin*, *142*(1), 79–106. https://doi.org/10.1037/bul0000026

Scherer, K. R., & Moors, A. (2019). The emotion process: Event appraisal and component differentiation. *Annual Review of Psychology*, *70*, 719–745. https://doi.org/10.1146/annurev-psych-122216-011854

Sorce, J. F., Emde, R. N., Campos, J. J., & Klinnert, M. D. (1985). Maternal emotional signaling: Its effect on the visual cliff behavior of 1-year-olds. *Developmental Psychology*, *21*(1), 195–200. https://doi.org/10.1037/0012-1649.21.1.195

Stussi, Y., & Pool, E. R. (2022). Multicomponential affective processes modulating food-seeking behaviors. *Current Opinion in Behavioral Sciences*, *48*, Article 101226. https://doi.org/10.1016/j.cobeha.2022.101226

Stussi, Y., Pourtois, G., Olsson, A., & Sander, D. (2021). Learning biases to angry and happy faces during Pavlovian aversive conditioning. *Emotion*, *21*(4), 742–756. https://doi.org/10.1037/emo0000733

Stussi, Y., Pourtois, G., & Sander, D. (2018). Enhanced Pavlovian aversive conditioning to positive emotional stimuli. *Journal of Experimental Psychology: General*, *147*(6), 905–923. https://doi.org/10.1037/xge0000424

Vogl, E., Pekrun, R., Murayama, K., & Loderer, K. (2020). Surprised-curious-confused: Epistemic emotions and knowledge exploration. *Emotion*, *20*(4), 625–641. https://doi.org/10.1037/emo0000578

Vollberg, M. C., & Sander, D. (2024). Hidden reward: Affect and its prediction errors as windows into subjective value. *Current Directions in Psychological Science*. Advance online publication. https://doi.org/10.1177/09637214231217678

Wuensch, L., Pool, E. R., & Sander, D. (2021). Individual differences in learning positive affective value. *Current Opinion in Behavioral Sciences*, *39*, 19–26. https://doi.org/10.1016/j.cobeha.2020.11.001

Zentall, T. R., & Galef, B. G. (Eds.). (1988). *Social learning: Psychological and biological perspectives*. Psychology Press.

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

Bayes beyond the predictive distribution

Anna Székely[a,b]* and Gergő Orbán[a]

[a]Department of Computational Sciences, HUN-REN Wigner Research Centre for Physics, Budapest, Hungary and [b]Department of Cognitive Science, Faculty of Natural Sciences, Budapest University of Technology and Economics, Budapest, Hungary
szekely.anna@wigner.hu
http://golab.wigner.mta.hu/people/anna-szekely/
orban.gergo@wigner.mta.hu
http://golab.wigner.mta.hu/people/gergo-orban/

*Corresponding author.

doi:10.1017/S0140525X24000086, e166

Abstract

Binz et al. argue that meta-learned models offer a new paradigm to study human cognition. Meta-learned models are proposed as alternatives to Bayesian models based on their capability to learn identical posterior predictive distributions. In our commentary, we highlight several arguments that reach beyond a predictive distribution-based comparison, offering new perspectives to evaluate the advantages of these modeling paradigms.

In their review, Binz et al. propose a framework for studying the adaptive nature of the mind. They propose that recent advances in machine learning empower meta-learning paradigms to be used as a flexible and general framework for studying the computations, the representations, and even the neuronal processes underlying learning. The authors put forward a number of arguments that provide support for such a paradigm. In this commentary, we aim to reflect on these arguments in order to better identify the advantages and limits of using meta-learned models instead of Bayesian ones.

The authors pit the meta-learning paradigm against Bayesian approaches. Bayesian models provide a similarly general framework for formulating learning problems as meta-learned models, but the two paradigms differ in the principles that guide model construction. In contrast with the primarily data-driven approach of meta-learned models, Bayesian approaches formulate the computational challenge humans face when performing task(s) through the definition of likelihood and priors, which summarize our assumptions about the relevant quantities of the computational challenge and our prior beliefs about these quantities. In other words, when constructing a Bayesian model, one needs to define a generative model of the task and also the relevant quantities that shape the learning procedure, which instantly provides a set of testable hypotheses and, thus, an opportunity to better understand cognition. The authors challenge the Bayesian approach by pointing out that in complex tasks, both defining and evaluating the likelihood can be impossible, and the function classes that Bayesian models rely on can be severely constrained. The authors argue that these challenges can be circumvented by using meta-learned models instead. To support the paradigm shift, the authors cite promising new studies that explore the equivalence of meta-learned models and Bayesian approaches. While these unifying views certainly contribute to a better understanding of learning, some aspects of these views deserve further consideration.

The authors argue that it is the posterior predictive distribution that a model ultimately learns, and thus, this quantity provides a platform to compare alternative approaches. The posterior predictive distribution is then used to establish the equivalence of Bayesian and meta-learned models. We would challenge this view based on two observations. First, it is important to point out that in its general form, the posterior predictive distribution is not a quantity that is invariant for a set of tasks, but it depends on the choice of the prior. This also means that the equivalence of the meta-learner and the Bayesian learner is constrained. This constraint can be illuminated by considering the contribution of the priors in Bayesian models. The effect of prior is most pronounced when data are scarce. In such cases, the equivalence is hard to establish as it is unclear what sort of prior the meta-learner model implicitly assumes. When data are abundant, however, the contribution of the prior diminishes, and in such cases, it is easier to establish the equivalence of the two model classes. Second, comparing Bayesian models and deep networks based on predictive performance alone ignores the power of having a framework that permits combining structured knowledge representations with powerful inference (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Kemp, Perfors, & Tenenbaum, 2007; Kemp & Tenenbaum, 2008; Tenenbaum, Griffiths, & Kemp, 2006, 2011). A key benefit of Bayesian modeling is the characterization of generative models that could plausibly account for the behavioral outcomes. Creating and testing hypotheses regarding these generative models enables us to better understand the computations that underlie cognition and give rise to the behavioral outcome.

The authors refer to inductive biases that can be transparently captured by meta-learned models, some of which are not necessarily easy to capture in Bayesian models. While we agree that some forms of inductive biases are readily delivered by these meta-learned models, Bayesian models too are capable of investigating relevant inductive biases. These inductive biases might include assumptions about the function classes that learning operates on (Kemp & Tenenbaum, 2008) or assumptions about the computational complexity of the generative model (Csikor, Meszéna, & Orbán, 2023) both of which can be phrased through the definition of the likelihood. Such inductive biases can be explored by pitting them against alternatives and assessing the models' power to predict human learning. In summary, we argue that characterization of learning through the specification of the generative model, comprised of the prior and the likelihood, makes it possible to explore the assumptions behind the models, which assumptions may remain hidden in meta-learned models.

Finally, it's important to clarify that we agree with the authors that more flexible tools provide unique opportunities to study a broader class of phenomena. However, recent advances in Bayesian models open new opportunities in this aspect, for example, variational autoencoders (Nagy, Török, & Orbán, 2020; Spens & Burgess, 2024), non-parametric methods (Éltető, Nemeth, Janacsek, & Dayan, 2022; Heald, Lengyel, & Wolpert, 2021; Török et al., 2022), or probabilistic programming (Lake, Salakhutdinov, & Tenenbaum, 2015), might leverage the need to meticulously define model architectures a priori by the experimenter and will complement the data-driven meta-learning approach proposed by the authors. In particular, the contribution of changing inductive biases to task performance in humans has been recently investigated in an implicit learning paradigm using a non-parametric Bayesian approach (Székely et al., 2024). In general, a combination of flexible nonlinear Bayesian models with structure learning is particularly appealing and has proven to be a valuable tool in continual learning (Achille et al., 2018; Rao et al., 2019).

Financial support. Supported by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory.

Competing interest. None.

References

Achille, A., Eccles, T., Matthey, L., Burgess, C. P., Watters, N., Lerchner, A., & Higgins, I. (2018). Life-long disentangled representation learning with cross-domain latent homologies. NeurIPS.

Csikor, F., Meszéna, B., & Orbán, G. (2023). Top-down perceptual inference shaping the activity of early visual cortex. *BioRxiv*. https://doi.org/10.1101/2023.11.29.569262

Éltető, N., Nemeth, D., Janacsek, K., & Dayan, P. (2022). Tracking human skill learning with a hierarchical Bayesian sequence model. *PLoS Computational Biology*, *18*(11), e1009866. https://doi.org/10.1371/journal.pcbi.1009866

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364. https://doi.org/10.1016/j.tics.2010.05.004

Heald, J. B., Lengyel, M., & Wolpert, D. M. (2021). Contextual inference underlies the learning of sensorimotor repertoires. *Nature*, *600*, 489–493. https://doi.org/10.1038/s41586-021-04129-3

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321. https://doi.org/10.1111/j.1467-7687.2007.00585.x

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *PNAS*, *105*(31), 10687–10692. https://doi.org/10.1073/pnas.0802631105

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338. Retrieved from www.sciencemag.org

Nagy, D. G., Török, B., & Orbán, G. (2020). Optimal forgetting: Semantic compression of episodic memories. *PLoS Computational Biology*, *16*(10), e1008367. https://doi.org/10.1371/journal.pcbi.1008367

Rao, D., Visin, F., Rusu, A. A., Teh, Y. W., Pascanu, R., & Hadsell, R. (2019). Continual unsupervised representation learning. NeurIPS.

Spens, E., & Burgess, N. (2024). A generative model of memory construction and consolidation. *Nature Human Behaviour*, *8*, 526–543. https://doi.org/10.1038/s41562-023-01799-z

Székely, A., Török, B., Kiss, M. M., Janacsek, K., Németh, D., & Orbán, G. (2024). Identifying transfer learning in the reshaping of inductive biases. *PsyArxiv*.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318. https://doi.org/10.1016/j.tics.2006.05.009

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Török, B., Nagy, D. G., Kiss, M., Janacsek, K., Németh, D., & Orbán, G. (2022). Tracking the contribution of inductive bias to individualised internal models. *PLoS Computational Biology*, *18*(6), e1010182. https://doi.org/10.1371/journal.pcbi.1010182

Meta-learning and the evolution of cognition

Walter Veit[a]* and Heather Browning[b]

[a]Department of Philosophy, University of Reading, Reading, UK and
[b]Department of Philosophy, University of Southampton, Southampton, UK
wrwveit@gmail.com
DrHeatherBrowning@gmail.com
https://walterveit.com/
https://www.heatherbrowning.net/

*Corresponding author.

doi:10.1017/S0140525X24000177, e167

Abstract

Meta-learning offers a promising framework to make sense of some parts of decision-making that have eluded satisfactory explanation. Here, we connect this research to work in animal behaviour and cognition in order to shed light on how and whether meta-learning could help us to understand the evolution of cognition.

Computational models of learning were historically largely designed by hand. But this has changed dramatically in the last decade with the rise of so-called meta-learning models that have their priors updated through feedback with the environment, thus offering a better approximation of how human cognition works due to the inclusion of flexibility and agency. Indeed, they have been able to explain some puzzling phenomena that had resisted satisfactory explanations. Yet, this wealth of research has not been unified into anything like a coherent account. The lack of such an account motivated Binz et al. to offer a synthesis of this research and framework for future research by drawing on Bayesian inference models of cognition and rational models of cognition (e.g., Anderson, 2013).

In this commentary, we do not aim to challenge their proposal, which we find very compelling. In synthesising the scattered literature and explaining meta-learning in an accessible manner, we believe the authors to be more than successful, and we share their optimism for applications of meta-learning models of cognition. Instead of criticising an aspect of their approach, we will here follow up on a question they themselves raise, but do not pursue further: "How much of it [meta-learning] is based on evolutionary or developmental processes?" (2024, p. 11). We hope to aid both the development of better meta-learning models as well as a better understanding of human learning by investigating the evolution of meta-learning from simple animals to humans. As Dennett (1995) once put it, natural selection is an acid that leaves nothing untouched, and meta-learning as we shall argue is no exception.

Binz et al. show their optimism for how meta-learning can help in understanding how cognition can develop in agents through repeated interactions with the environment, which can provide a useful model to understand human developmental processes, though they admit more research would be needed. But more interestingly perhaps – and not surprising to anyone who emphasises that development often recapitulates evolutionary processes – there is also the potential to use meta-learning models to help us understand the evolution of cognition more generally. Binz et al. urge us to consider the more complex tasks we find in natural settings for humans, but that point is worth extending towards non-human animals. While they note that meta-learning models may help us to bridge the two traditions of connectionism and Bayesian learning, an evolutionary perspective could help us to merge these traditions.

If we ask why cognition evolved – or here more specifically why creatures may have evolved meta-learning capacities – we can draw on the aforementioned puzzling tasks that meta-learning helps us to explain, such heuristic-based decision-making, as some of the authors have noted elsewhere (Binz, Gershman, Schulz, & Endres, 2022). Non-human animals, after all, also use heuristic strategies to navigate their environments. Admittedly, animal models of behaviour typically satisfy themselves with "hand-designed" algorithms of behaviour, but such models are deliberately simple to account for trade-offs between particular considerations, for example, optimal foraging under conditions of high predator-density. Studies of animal cognition have already established that animals can solve more complex problems than was predicted (Andrews & Monsó, 2021). When Binz et al. describe the four advantages a meta-learning model has over a standard Bayesian model, two key features emerge that are highly relevant to an evolutionary account of meta-learning: Resource limitations, and the lack of prior information about the environment. When considering both of these features and the way they operate on organisms in the wild, the ecological plausibility of even an early evolution of meta-learning capacities becomes quite plausible.

Meta-learning models are able to limit the complexity of the algorithms they use, to reduce strain on resources. Under the constraints of natural selection, resource limitations play a strong role in determining the optimal strategy for organismal behaviour and/or phenotype. Out in the world, organisms will have constraints on brain size (and subsequently, memory capacity and processing power), as well as the time and energy availability for running cognitive computations. A system that provides a method for limiting the complexity of more difficult algorithms – as meta-learning does – will therefore have a strong advantage for organisms operating under normal constraints.

It is also a common feature of the environments in which animals find themselves that they will lack prior information about these environments (Veit, 2023). For any animal that lives in a variable or changing environment, or those with a complex and flexible behavioural repertoire, they cannot know in advance what they will encounter throughout their lifetimes, such as the distributions of the types of functions they will come across. A learning model that allows an organism to improve its learning over repeated encounters with and sampling of its environment will be selectively advantageous in these contexts as they can adapt to whatever circumstances in which they find themselves. Conversely, animals who evolve and develop within stable environments with a fairly fixed set of challenges may do better with pre-set learning algorithms that are optimised for this environment, to avoid the complexity and time investment required for meta-learning.

A meta-learning perspective on the evolution of animal cognition also fits with our current neuroscientific knowledge on cognitive architectures, as well as the empirical data on animal learning. For instance, Binz et al. note that many species (including humans) have been shown to improve their learning strategies over time. This empirical evidence supports the evolutionary story we have sketched here. Unfortunately, much work remains to be done in order to understand the evolution of cognition, but we hope to have successfully shown that meta-learning could offer a promising framework for enhancing such understanding, due to its inherent link to the adaptive agency of living systems.

Financial support. No funding to report.

Competing interest. None.

References

Anderson, J. R. (2013). *The adaptive character of thought*. Psychology Press.

Andrews, K., & Monsó, S. (2021). Animal cognition. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/spr2021/entries/cognition-animal/

Binz, M., Gershman, S. J., Schulz, E., & Endres, D. (2022). Heuristics from bounded meta-learned inference. *Psychological Review*, *129*, 1042–1077. https://doi.org/10.1037/rev0000330

Dennett, D. C. (1995). *Darwin's dangerous idea: Evolution and the meaning of life*. Simon and Schuster.

Veit, W. (2023). *A philosophy for the science of animal consciousness*. Routledge.

The reinforcement metalearner as a biologically plausible meta-learning framework

Tim Vriens[a], Mattias Horan[b], Jacqueline Gottlieb[c,d]* and Massimo Silvetti[a]

[a]Institute of Cognitive Sciences and Technologies, CNR, Rome, Italy; [b]Sainsbury Wellcome Centre, University College London, London, UK; [c]Department of Neuroscience, Columbia University, New York, NY, USA and [d]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA
Tim.Vriens@unicampus.it,
mattias.horan.19@ucl.ac.uk,
jg2141@columbia.edu,
massimo.silvetti@istc.cnr.it
https://zuckermaninstitute.columbia.edu/jacqueline-gottlieb-phd
https://ctnlab.it/index.php/massimo-silvetti/, https://www.istc.cnr.it/en/people/massimo-silvetti

*Corresponding author.

doi:10.1017/S0140525X24000219, e168

Abstract

We argue that the type of meta-learning proposed by Binz et al. generates models with low interpretability and falsifiability that have limited usefulness for neuroscience research. An alternative approach to meta-learning based on hyperparameter optimization obviates these concerns and can generate empirically testable hypotheses of biological computations.

Binz et al. describe four different meta-learning approaches and focus on the last one – methods for learning arbitrary new tasks without the need for a priori hypotheses about brain or cognitive architectures. They show that this approach can be implemented in recurrent neural networks (RNNs) that are universal approximators (Hornik, Stinchcombe, & White, 1989), and argue that it is powerful in producing Bayesian (near-optimal) learning in an arbitrarily large set of cognitive tasks. While acknowledging the power of the proposed framework for artificial intelligence (AI), we question its usefulness in cognitive and neuroscience research. We argue that an alternative approach of hyperparameter optimization (which was first proposed by Doya, 2002, and is mentioned but not discussed by Binz et al.) is far more powerful for this role.

To be valuable for empirical research, a computational framework should generate models that are interpretable in neurocognitive terms and make predictions that can be falsified or confirmed through empirical tests. The internal computations used by the models should be analogous to those of neurocognitive systems (e.g., attention, memory, valuation, etc.; e.g., Castelvecchi, 2016), and predict activity patterns that can be empirically validated. The framework advocated by Binz et al. has neither property, and instead generates models that are governed by immense numbers of free parameters (up to billions) and are not interpretable in cognitive terms, amounting to a "black box" data-driven approach.

A hyperparameter optimization approach alleviates these concerns by constraining the models it generate to emulate biologically plausible architectures. This allows for formulating and testing mechanistic hypotheses that are based in established literature. The reinforcement meta-learner (RML) model is a good illustration of this framework in the context of executive function (Silvetti, Vassena, Abrahamse, & Verguts, 2018).

Consistent with abundant empirical evidence on biological executive circuits (e.g., Shackman et al., 2011; Silvetti, Seurinck, van Bochove, & Verguts, 2013; Varazzani, San-Galli, Gilardeau, & Bouret, 2015; Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011), the RML emulates interactions between the medial prefrontal cortex (MPFC) and two catecholamine nuclei – the ventral tegmental area, releasing dopamine (DA), and the locus coeruleus, releasing norepinephrine (NE). The MPFC module monitors reward rates conveyed by DA and, when detecting a "need for control" (e.g., a decrease in the rates), calls for the release of NE and DA. In turn, these neurotransmitters are broadcast to task-specific cognitive modules and enhance their efficiency, thereby restoring performance and reward rates. The MPFC registers a boost of neurotransmitter release as a cost and uses Bayesian and reinforcement-learning (RL) optimization to learn control settings that maximize rewards while minimizing costs. The RML thus uses traditional Bayesian/RL optimization frameworks to simultaneously regulate motor input and internal cognitive computations, thus modeling both first-order performance and its executive (meta-level) control.

Recent studies have shown that the RML explains empirical findings that have long stumped traditional frameworks, including nonstandard reward modulations in visual areas (Horan, Daddaoua, & Gottlieb, 2019; Silvetti, Lasaponara, Daddaoua, Horan, & Gottlieb, 2023) and curiosity – the intrinsic desire to obtain information in the absence of instrumental rewards (Daddaoua, Lopes, & Gottlieb, 2016; Horan et al., 2019; Silvetti et al., 2023). By monitoring the volatility of the environment, the RML provides a meta-learning-based explanation of the empirical finding of volatility-sensitive learning rates (Silvetti et al., 2013, 2018). Moreover, when coupled to modules emulating memory, motor output, decision making, or attention, the RML reproduces a wide array of behavioral and neural results related, respectively, to memory capacity, motor effort, adaptive regulation of learning rates, and instrumental or curiosity-driven information gathering (Silvetti et al., 2018, 2023). Thus, despite its biologically constrained architecture, the RML gains considerable flexibility and generalizability because it can control different task-specific cognitive computations.

Because the RML uses a biologically plausible architecture with a parsimonious parameter set, it generates a rich set of novel predictions that can be tested against empirical data. These predictions involve possible relationships between behavior and neural activity, between neural activity and neurotransmitter release, and between activity in different brain structures. Existing versions of the RML make predictions about individual computations (e.g., how much memory effort to engage in a particular context) but future versions can be extended to probe how the brain arbitrates between computations (e.g., how it trades-off between relying on memory versus acquiring new sensory information when performing a task).

In conclusion, different meta-learning approaches can differ greatly in their comparative strengths. The entirely unconstrained approach discussed by Binz et al. may be desirable for AI applications where there is no need for biological constraints, for example, when developing an algorithm for a self-driving car, or optimizing planning in multiple tasks. In contrast, we believe that a biologically constrained meta-learning framework is vastly superior for advancing cognitive and neuroscience research (Marblestone, Wayne, & Kording, 2017). Such a biologically constrained framework is grounded in the neuroscientific literature, and can generate testable and falsifiable hypotheses about neurobiological processes underlying cognitive function.

Acknowledgments. Tim Vriens is a PhD student enrolled in the National PhD in Artificial Intelligence, XXXVII cycle, course on Health and life sciences, hosted by Università Campus Bio-Medico di Roma, Italy.

Financial support. M. S. is funded by the Italian Ministry of University and Research, PRIN 2022 program, Grant No. 64.20227MPSEH. M. H. is supported via the Sainsbury Wellcome Centre PhD Programme and has received grants from Reinholdt W. Jorck og Hustrus Fond, Knud Højgaards Fond and Anglo-Danish Society

Competing interests. None.

References

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, *538*, 20–23. https://doi.org/10.1038/538020a

Daddaoua, N., Lopes, M., & Gottlieb, J. (2016). Intrinsically motivated oculomotor exploration guided by uncertainty reduction and conditioned reinforcement in non-human primates. *Scientific Reports*, *6*(1), Article 1. https://doi.org/10.1038/srep20202

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*(4–6), 495–506. https://doi.org/10.1016/s0893-6080(02)00044-8

Horan, M., Daddaoua, N., & Gottlieb, J. (2019). Parietal neurons encode information sampling based on decision uncertainty. *Nature Neuroscience*, *22*(8), 1327–1335. https://doi.org/10.1038/s41593-019-0440-1

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Marblestone, A. H., Wayne, G., & Kording, K. P. (2017). Understand the cogs to understand cognition. *Behavioral and Brain Sciences*, *40*, e272. https://doi.org/10.1017/S0140525X17000218

Shackman, A. J., Salomons, T. V., Slagter, H. A., Fox, A. S., Winter, J. J., & Davidson, R. J. (2011). The integration of negative affect, pain, and cognitive control in the cingulate cortex. *Nature Reviews. Neuroscience*, *12*(3), 154–167. https://doi.org/10.1038/nrn2994

Silvetti, M., Lasaponara, S., Daddaoua, N., Horan, M., & Gottlieb, J. (2023). A reinforcement meta-learning framework of executive function and information demand. *Neural Networks*, *157*, 103–113. https://doi.org/10.1016/j.neunet.2022.10.004

Silvetti, M., Seurinck, R., van Bochove, M., & Verguts, T. (2013). The influence of the noradrenergic system on optimal control of neural plasticity. *Frontiers in Behavioral Neuroscience*, *7*, 160. https://www.frontiersin.org/articles/10.3389/fnbeh.2013.00160

Silvetti, M., Vassena, E., Abrahamse, E., & Verguts, T. (2018). Dorsal anterior cingulate-brainstem ensemble as a reinforcement meta-learner. *PLoS Computational Biology*, *14*(8), e1006370. https://doi.org/10.1371/journal.pcbi.1006370

Varazzani, C., San-Galli, A., Gilardeau, S., & Bouret, S. (2015). Noradrenaline and dopamine neurons in the reward/effort trade-off: A direct electrophysiological comparison in behaving monkeys. *The Journal of Neuroscience*, *35*(20), 7866–7877. https://doi.org/10.1523/JNEUROSCI.0454-15.2015

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670. https://doi.org/10.1038/nmeth.1635

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

Integrative learning in the lens of meta-learned models of cognition: Impacts on animal and human learning outcomes

Bin Yin*, Xi-Dan Xiao, Xiao-Rui Wu and Rong Lian

School of Psychology, Fujian Normal University, Fuzhou, China.
byin@fjnu.edu.cn
Xiaoxid8899@foxmail.com
Wuxiaorui520@hotmail.com
Lianrong1122@126.com

*Corresponding author.

doi:10.1017/S0140525X2400027X, e169

Abstract

This commentary examines the synergy between meta-learned models of cognition and integrative learning in enhancing animal and human learning outcomes. It highlights three integrative learning modes – holistic integration of parts, top-down reasoning, and generalization with in-depth analysis – and their alignment with meta-learned models of cognition. This convergence promises significant advances in educational practices, artificial intelligence, and cognitive neuroscience, offering a novel perspective on learning and cognition.

Binz et al.'s seminal paper on "Meta-Learned Models of Cognition" offers a transformative view on cognitive modeling, shifting the traditional paradigm toward a more dynamic and experience-based approach. The authors convincingly argue for the superiority of meta-learned models in acquiring inductive biases from experience, as opposed to the rigid, hand-designed structures of traditional models like cognitive architectures and Bayesian models of cognition. This shift represents not only a theoretical advancement but also a practical one, providing a more realistic and adaptable framework for understanding cognitive processes.

Crucially, the paper's synthesis of meta-learning with rational analysis presents an exciting pathway for constructing Bayes-optimal learning algorithms. This approach resonates strongly with integrative learning theories that we have been working on, suggesting a shared trajectory toward developing learning models that can adapt and thrive amidst complexity. Integrative learning refers to the cognitive process of actively integrating learning materials under the influence of metacognition, resulting in an efficient and profound understanding and mastery of knowledge (Lian, 2018). It represents a psychological learning process where metacognition and cognition are highly unified (Yin, Wu, & Lian, 2020, 2023). This learning process model encompasses three modes: holistic integration of parts, top-down reasoning, and generalization for in-depth analysis (Rong Lian, personal communication, March 2020).

Firstly, "holistic integration of parts" involves learners first grasping the overall concept of the subject, establishing a comprehensive initial understanding. This is followed by an exploration of specific parts of the material, with each part being connected back and integrated into this broader understanding, thus reinforcing and enriching the overall comprehension. This "whole-part-whole" learning process has been shown to play a positive role not only in animal learning (Yin et al., 2020, 2023) but also in human learning processes. In studies with university students learning online network knowledge, it was found that, compared to a non-integrative learning group, the integrative learning group better synthesized and processed fragmented online knowledge, resulting in superior learning performance (Huang, 2021). This indicates that individuals, during the learning process, activate metacognition which combines prior knowledge and experience to adjust and optimize new knowledge within working memory, thereby enhancing online learning effectiveness (Mayer, 1997).

Secondly, "top-down reasoning" emphasizes beginning with more broad and generalized high-level concepts and systematically mastering more specific lower-level knowledge points. By understanding and applying higher-level knowledge, they deduce and explain lower-level knowledge, thereby forming a clear, logically structured knowledge framework. Lan (2022) explored the impact of this approach on learning outcomes using a study-recognition paradigm. The results showed that compared to bottom-up learning, top-down reasoning enabled learners to use attention resources more reasonably and effectively, facilitating the relational processing and integration of specific items. Additionally, cognitive semantic processing was smoother, and the integration difficulty between semantics was reduced, significantly enhancing memory effects. Event-related potential (ERP) technology was used to explore the underlying neural mechanisms, revealing that the top-down reasoning group had larger N1 amplitudes and significantly smaller N400 than the bottom-up learning group when learning specific examples. This indicates that top-down reasoning learners more effectively utilized higher-level knowledge, focusing attention resources on more organized processing of specific examples. From the perspective of semantic priming effects, once semantic concepts in memory are activated, their activation can spread to related nodes, increasing their activation level (Meyer & Schvaneveldt, 1971), making the learning of related target stimuli easier (Bueno & Frenck-Mestre, 2002).

Lastly, "generalization for in-depth analysis" starts with learners forming a general representation of the subject, laying the groundwork for the overall framework. Learners then delve into detailed components, engaging in thorough analysis and research. This in-depth study not only deepens understanding of each part but also enhances and refines the initial general framework, culminating in a multifaceted and detailed overall cognition. Chen (2023) explored the impact of this learning mode on university students' reading of texts of varying difficulty. Results showed that the "generalization and in-depth analysis" group had significantly higher understanding and memory scores under different text difficulty conditions compared to the control group, and also had a lower rate of knowledge forgetting. Reading is a process involving simultaneous extraction and construction of meaning (García Madruga et al., 2013). In reading, learners must use complex meaning construction processing to form a complete representation, where cognitive control plays a key role in focusing and switching attention, activating and updating representations (Wu, Tian, Chen, Chen, & Wang, 2021). The "generalization for in-depth analysis" mode helps learners construct complete representations more quickly, and through in-depth analysis and summarization of new knowledge, continuously adjust and optimize their meaning representations, thereby promoting more refined processing and encoding of information. This process helps learners more effectively retrieve and remember encoded information.

The three modes of integrative learning—holistic integration of parts, top-down reasoning, and generalization with in-depth analysis—closely align with the computational processes of meta-learned models of cognition. This alignment is pivotal, as the initial encounter with a subject in a holistic, higher-level, or generalized manner is essential for setting effective starting metaparameters that guide the learning process. Such an encounter provides a foundational understanding from which learners can refine their perceptions and strategies in a targeted manner. Within this adaptive framework, learners then engage in a cyclical process of interaction, analysis, and metacognitive adjustment, fine-tuning their approach based on this foundational overview and their evolving comprehension of the subject matter. This methodology not only embodies the adaptability characteristic of meta-learning but also supports real-time adjustments during the learning process. As a result, it leads to learning outcomes that are both precisely tailored to the learner's needs and highly effective. Consequently, emphasizing the value of an initial holistic overview reinforces the importance of integrative learning within the meta-learning paradigm. It empowers learners to dynamically adjust their information processing strategies from the outset, significantly enhancing and adapting their learning experiences to achieve optimal outcomes (Rabinowitz, 2019).

The implications of this area of research are vast, offering new directions for educational practices, artificial intelligence, and cognitive neuroscience. In education, these insights could lead to more personalized and effective learning strategies, tailored to individual (meta-)cognitive patterns. For AI, integrating these models could result in more adaptive and intuitive systems, better mimicking human learning processes. In cognitive neuroscience, this research offers potential for deeper understanding of brain-based learning mechanisms. Altogether, this represents a significant stride in our comprehension and application of cognitive and learning sciences, opening new avenues for exploration and innovation.

Financial support. This work was supported by the Humanities and Social Science Funds of the Ministry of Education of China, Project No. 23YJAZH183.

Competing interest. None.

References

Bueno, S., & Frenck-Mestre, C. (2002). Rapid activation of the lexicon: A further investigation with behavioral and computational results. *Brain and Language*, *81*(1–3), 120–130. http://doi.org/10.1006/brln.2001.2511

Chen, S. (2023). *The effect of integrative learning on text reading in elementary and university students* (Master's thesis). Fujian Normal University.

García Madruga, J. A., Elosúa, M. R., Gil, L., Gómez Veiga, I., Vila, J. Ó., Orjales, I., … Duque, G. (2013). Reading comprehension and working memory's executive processes: An intervention study in primary school students. *Reading Research Quarterly*, *48*(2), 155–174. http://doi.org/10.1002/rrq.44

Huang, J. (2021). *The influence of integrative learning on recognition and retrieval of image and text* (Master's thesis). Fujian Normal University. http://doi.org/10.27019/d.cnki.gfjsu.2021.000798

Lan, R. (2022). *The memory advantage effect of top-down reasoning and its cognitive neural mechanism* (Master's thesis). Fujian Normal University. https://doi.org/10.27019/d.cnki.gfjsu.2022.000519

Lian, R. (2018). Integrative learning: Exploring new ways of learning. Fujian Provincial Learning Science Society Symposium, Fuzhou, China.

Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychologist*, *32*(1), 1–19. https://doi.org/10.1207/s15326985ep3201_1

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234. http://dx.doi.org/10.1037/h0031564

Rabinowitz, N. C. (2019). Meta-learners' learning dynamics are unlike learner'. arXiv preprint arXiv:1905.01320. https://doi.org/10.48550/arXiv.1905.01320

Wu, S., Tian, L., Chen, J., Chen, G., & Wang, J. (2021). Exploring the cognitive mechanism of irrelevant speech effect in Chinese reading: Evidence from eye movements. *Acta Psychologica Sinica*, *53*(7), 729–745. http://dx.doi.org/10.3724/SP.J.1041.2021.00729

Yin, B., Wu, X., & Lian, R. (2020). An animal behavioral model for the concept of "Integrative Learning". *Acta Psychologica Sinica*, *52*(11), 1278–1287. https://doi.org/10.3724/SP.J.1041.2020.01278

Yin, B., Wu, X.-R., & Lian, R. (2023). "Integrative learning" promotes learning but not memory in older rats. *PeerJ*, *11*, e15101. http://doi.org/10.7717/peerj.15101

Authors' Response

Meta-learning: Data, architecture, and both

Marcel Binz[a,b]*, Ishita Dasgupta[c], Akshay Jagadish[a,b], Matthew Botvinick[c], Jane X. Wang[c] and Eric Schulz[a,b]

[a]Max Planck Institute for Biological Cybernetics, Tübingen, Germany; [b]Helmholtz Institute for Human-Centered AI, Munich, Germany and [c]Google DeepMind, London, UK
marcel.binz@helmholtz-munich.de
idg@google.com
akshay.jagadish@helmholtz-munich.de
botvinick@google.com
wangjane@google.com
eric.schulz@helmholtz-munich.de

*Corresponding author.

doi:10.1017/S0140525X24000311, e170

Abstract

We are encouraged by the many positive commentaries on our target article. In this response, we recapitulate some of the points raised and identify synergies between them. We have arranged our response based on the tension between data and architecture that arises in the meta-learning framework. We additionally provide a short discussion that touches upon connections to foundation models.

R1. Introduction

In our target article, we sketched out a research program around the idea of meta-learned models of cognition. The cornerstone of this research program was the observation that neural networks, such as recurrent neural networks, can be trained via meta-learning to mimic Bayesian inference without being explicitly designed to do so (Ortega et al., 2019). This positions the resulting meta-learned models ideally for applications in the context of rational analyses of cognition (Anderson, 2013). Yet, meta-learning additionally enables us to do things that are not possible with other existing methods, thereby pushing the boundaries of rational analyses. Not only is the framework built on solid theoretical grounds, but it also enjoys growing empirical support. Meta-learned models account for a wide range of phenomena that pose a challenge to traditional models, such as the ability for compositional reasoning (Jagadish, Binz, Saanum, Wang, & Schulz, 2023; Lake & Baroni, 2023) or the reliance on heuristic strategies (Binz, Gershman, Schulz, & Endres, 2022; Dasgupta, Schulz, Tenenbaum, & Gershman, 2020).

We believe this research direction is particularly exciting because it allows us to reconceptualize different cognitive processes, including learning, planning, reasoning, and decision-making, into one unified process: The forward dynamics of a deep neural network. In the terminology of modern large language models (LLMs), this ability to acquire knowledge via a simple forward pass is also known as in-context learning (Brown et al., 2020). In-context learning stands in contrast to traditional means of knowledge acquisition in neural networks that requires weight adjustment via gradient descent (hence, it is referred to as in-weights learning). Indeed, there are close connections between meta-learning and training LLMs to which we will return at the end of our response.

Many commentators shared our excitement about this new technology. **McCoy & Griffiths** write that the “direction laid out by Binz et al. is exciting.” **Pesnot-Lerousseau & Summerfield** suggest that “this computational approach provides an interesting candidate solution for some of nature’s most startling and puzzling behaviours” and that “meta-learning is a general theory of natural intelligence that is – more than [its] classical counterpart – fit for the real world.” **Grant** says that “it is an exciting time to be working with and on meta-learning toolkit” but also points out that “many aspects remain open.” We agree with this sentiment: Meta-learning is a powerful framework that provides us with the toolkit to build candidate theories of human cognition, but we still must figure out the details and precise instantiations that best describe it.

In the target article, we have framed our argument from a Bayesian angle. Although this offers an invaluable perspective, it does somewhat undermine the role that neural networks play in this context. Indeed, it is really the marriage between Bayesian and neural network models that gives meta-learning its power. There were several commentators who picked up on this. We agree with **Ong, Zhi-Xuan, Tenenbaum, & Goodman (Ong et al.)** who “suggest that the meta-learning approach could be further strengthened by considering connectionist and Bayesian approaches, rather than exclusively one or the other.” **McCoy & Griffiths** perhaps put it best by saying that meta-learning “expand[s] the applicability of Bayesian approaches by reconciling them with connectionist models – thereby bringing together two successful research traditions that have often been framed as antagonistic.” This integration of research traditions is what enables us to build “constraints from experimental neuroscience, and ecologically relevant environments” into rational theories as suggested by **Grant**, thereby leading to more faithful and naturalistic models.

Many commentators also noted that meta-learning finds applications beyond studying standard human cognition. For example, **Fields & Glazebrook** suggest studying meta-learning “in more tractable experimental systems in which the implementing architecture can be manipulated biochemically and bioelectrically”, whereas **Veit & Browning** highlight that “there is also the potential to use meta-learning models to help us understand the evolution of cognition more generally.” **Nussenbaum & Hartley** furthermore point out that “these models are particularly useful for testing hypotheses about why learning processes change across development” because they allow us to arbitrate whether changes in an individual are due to an adaption to the external environment (i.e., changes in data) or to internal changes in cognitive capacity (i.e., changes in architecture). We are excited by these research directions as well.

We found the tension between data and architecture laid out by **Nussenbaum & Hartley** very useful, and have therefore decided to organize our response around it. We begin by discussing the commentaries that placed a focus on the importance of data for understanding human cognition (sect. R2), followed by those that focused on the importance of model architecture instead (sect. R3). The point where those two concepts meet will be the centerpiece of our discussion (sect. R4). We finish our response by clarifying some of the misunderstandings that have arisen from our original target article (sect. R5), before providing a general conclusion (sect. R6).

R2. Data matters more than we thought

Historically, cognitive models have been largely based on symbolic representations. Examples include models of heuristic decision-making, problem-solving, or planning. This modeling tradition is largely based on the premise that model architectures are the driving factor in determining behavior. Proponents of this approach often argue that symbolic representations are necessary to capture core ingredients of human cognition, such as decision-making, problem-solving, or planning (Marcus, 1998). The advent of Bayesian models of cognition expanded on this picture. Even though most Bayesian models are also based on symbolic representations, they are sensitive to the data that are expected to be encountered. If assumptions about the environment change, the behavior of these models changes as well. The past 30 years have shown that people are indeed adaptive to the environment, thereby providing considerable support to Bayesian models of cognition (Griffiths, Kemp, & Tenenbaum, 2008).

In contrast to models with symbolic representations, neural networks are based on distributed vector representations. Many have argued that neural networks are inherently ill-equipped for reasoning, planning, and problem-solving because they lack the symbolic representations of their cousins. Indeed, there is a whole line of research (known as neurosymbolic AI) attempting to fix these issues by incorporating symbolic processes into neural network architectures (De Raedt, Manhaeve, Dumancic, Demeester, & Kimmig, 2019). The framework of meta-learning demonstrates that this may not be necessary. It instead offers a proof-of-concept showing that – when trained on the right data – neural networks can exhibit many emergent phenomena that have traditionally been attributed to symbolic models, such as the ability for model-based (Wang et al., 2016) and compositional reasoning (Lake & Baroni, 2023). For example, as already discussed in our target article, Lake and Baroni (2023) have shown that a vanilla transformer architecture can be taught to make compositional inferences via meta-learning. Findings like this allow us to interpret human compositionality as “an emergent property of an inner-loop, in-context learning algorithm that is itself meta-learned” as discussed by **Russin, McGrath, Pavlick, & Frank**. Likewise, Wang et al. (2016) have shown that a simple meta-learned recurrent neural network can act like a model-based reinforcement learning algorithm, even though it does not contain any explicit architecture that facilitates model-based reasoning. The implications of these findings are vast as pointed out by **Pesnot-Lerousseau & Summerfield** who suggest that "many supposedly 'model-based' behaviours may be better explained by meta-learning than by classical models" and that meta-learning "invites us to revisit our neural theories of problem solving and goal-directed planning."

Taken together, this suggests that model architecture may not be as important as once thought for building systems with human-like reasoning capabilities. What matters much more than we initially thought however are the data these systems are trained on. If the data are generated by symbolic processes, meta-learning will pick up on this and compile these processes into the resulting models.

There is evidence from recent work in NeuroAI supporting the idea that data trumps architecture. In a large-scale analysis, Conwell, Prince, Kay, Alvarez, and Konkle (2022), for example, found different model architectures achieve near equivalent degrees of brain predictivity in the human ventral visual system and that the data they were trained on had a much bigger influence. Muttenthaler, Dippel, Linhardt, Vandermeulen, and Kornblith (2022) presented similar findings, suggesting that "model scale and architecture have essentially no effect on the alignment [between the representations learned by neural networks and] human behavioral responses, whereas the training dataset and objective function both have a much larger impact."

That puts the focus on the question of "what to learn?" **Prat & Lamm** argue that this is the hard problem in natural and artificial intelligence. They further point out that nature solves this problem via evolution and that we cannot handcraft the utility function (or error measurements) for each task separately. We sympathize with this perspective. However, the evolutionary perspective is not the most useful one when the goal is to build models of human cognition. We certainly do not want to simulate the entire process of evolution for this. If we want to avoid this, what can we do as an alternative? For one, we can use automated tools, such as Gibbs sampling with people (Harrison et al., 2020), that measure the priors and utility functions of people and plug the resulting data into our pipelines. That this is possible in the meta-learning framework has been recently demonstrated by Kumar et al. (2022). There is also recent work suggesting that the generation of data reflecting the real world can be automated using foundation models. Jagadish, Coda-Forno, Thalmann, Schulz, and Binz (2024) have, for example, shown that this is a promising approach for building models that can acquire human-like priors when trained on ecologically valid problems. In particular, they queried a LLM to generate naturalistic classification problems, trained a meta-learning system on these problems, and demonstrated that the resulting meta-learned models explain many effects observed in the literature.

R3. Yet, architecture matters too

However, it is certainly not only data that matter for understanding human cognition. Model architecture will still play a role as pointed out in several commentaries (e.g., **Schilling, Ritter, & Ohl**). In fact, there are already results showing that this is the case in the meta-learning setting. For example, Chan et al. (2022) studied the trade-off between in-context and in-weights learning and found that in-context learning only emerges when the training data exhibit certain data distributional properties. Importantly, this was only true in transformer-based models but not in recurrent models (which relied on in-weights learning instead). This demonstrates that different model architectures can lead to characteristically different behaviors, thereby highlighting that architecture *is* crucial – at least to some extent. From a cognitive perspective, the interesting question will be how much architecture is needed.

Many commentaries suggested that enhancing the black-box meta-learning framework with process-level structures will help us to better understand human cognition. We agree that this is an intriguing line of thought (see the discussion in our target article within sects. 2.4 and 5). In many cases, the commentaries added an additional dimension to our original proposal. We discuss some of these proposals in the following and place them into the context of our framework.

Sanborn, Yan, & Tsvetkov (Sanborn et al.) highlight that people often deviate from normative behavior (whether Bayesian or meta-learned). Earlier work has shown that many of these deviations can be captured by rational process models, which approximate posterior predictive distributions using posterior mean, posterior median, or other summary statistics. These rational process models play hand-in-hand with the meta-learning framework as pointed out by Sanborn et al.. Essentially, their proposal is to have a rational process model reason based on the meta-learned posterior predictive distribution. This combination brings together the best of both worlds as one does not even have to retrain the meta-learning model as in other approaches that induce limited computational resources into meta-learned models (Binz & Schulz, 2022; Saanum, Éltető, Dayan, Binz, & Schulz, 2023). That can be convenient from a practical perspective. We agree that this is an appealing property and look forward to how the interaction between these two frameworks will play out in the future.

In a similar vein, **Grant** argues that the meta-learning toolkit needs stronger architectural constraints. Her proposal emphasizes a connectionist implementation of meta-learning called model-agnostic meta-learning (MAML). MAML implements its stepwise updating using gradient descent as opposed to the models we focused on in our target article whose updating is implemented using neural network forward dynamics (which is also referred to as memory-based meta-learning). Although MAML involves meta-learning with the same objective as discussed in our target article, it differs in what is being meta-learned. In MAML, one adapts the initial weights of a neural network, such that subsequent gradient steps lead to optimal learning. That leads to an interesting class of gradient-based meta-learned models that have many (but not all) of the advantages discussed in our target article. MAML's key feature is that it allows for a seamless link between the algorithmic and the computational level of analysis. Future research should compare different classes of models against each other and find the one that best explains human behavior. It will be particularly exciting to pit gradient-based models (such as MAML) against memory-based models (including recurrent neural networks) and see which class of theories offers a better account of human behavior. Doing so will allow us to answer some of the big, outstanding questions of cognitive psychology and neuroscience, such as whether we can find any evidence for computations like gradient descent and backpropagation in the brain.

Last but not least, **Cea** and **Stussi, Dukes, & Sander** suggest that the meta-learning framework could benefit from the inclusion of affective elements. We agree that doing so can provide added value to the meta-learning research agenda. Yet, at the same time, this proposal highlights one of the tensions involved

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

in building complex systems, namely deciding on what should be prewired and what should be given the chance to emerge instead. To illustrate this point, let us consider one of the examples provided by Stussi et al. highlighting the importance of affective processes: For humans, positively valenced prediction errors are generally associated with a higher learning rate than negatively valenced prediction errors (Palminteri & Lebreton, 2022). In a recent study, we found that this characteristic emerges naturally in meta-learned models (Schubert, Jagadish, Binz, & Schulz, 2024), thereby illustrating that at least some affective processes are already present in meta-learned models.

Ultimately, determining which inductive biases should be prewired and which should be learned from data depends on which research question one is investigating. If one wants to obtain a process-level understanding of a phenomenon, there is no better way than formalizing that phenomenon mathematically and simulating it in silico. If the goal, on the other hand, is to simply induce superhuman-like general abilities in a computational model, modern machine learning research, such as the work on AlphaZero, has taught us that we should keep the amount of prewiring limited and instead rely mainly on the data itself (Sutton, 2019).

R4. Transcending levels of analysis

The full power of meta-learning does not solely come from its close ties to Bayesian inference – the algorithmic implementation also matters. To get a more complete understanding of human cognition, it seems likely that we need to consider both data and architecture. Meta-learning allows us to do this by bringing together two modeling traditions that have focused on these two aspects individually. It combines the advantages of Bayesian models – which feature powerful, data-dependent inductive biases – and neural network models – which come with a vast amount of architectural design choices – seamlessly. To quote Ortega (2020), meta-learning "brings back Bayesian statistics within deep learning without even trying – no latents, no special architecture, no special cost function, nada."

That was also recognized by some of the commentators. **McCoy & Griffiths** state that meta-learning "reconcil[es Bayesian approaches] with connectionist models – thereby bringing together two successful research traditions that have often been framed as antagonistic". This feature allows the framework to effortlessly jump between different levels of analysis, from the computational over the algorithmic to the implementational. Furthermore, although neural networks have been often criticized for not being able to engage in symbolic reasoning, meta-learning illustrates that it is, in principle, possible to equip neural networks with symbolic inductive biases.

Nussenbaum & Hartley highlight potential applications in the context of developmental psychology. Here, one of the central questions involves identifying whether "age-related changes in learning reflect adaptation to age-varying 'external' ecological problems *or* 'internal' changes in cognitive capacity." We believe that this is an exciting research direction. In fact, we have recently applied some of these ideas to test whether the developmental trajectories of children in the context of intuitive physics can be captured with deep generative models by manipulating the amount of training data or the system's computational resources (Buschoff, Schulz, & Binz, 2023).

However, the strict dichotomy between data and architecture is likely to be false. Instead, the two interact with each other over the lifespan, as also pointed out by **Nussenbaum & Hartley**. Meta-learning allows us to disentangle the two and study them jointly or separately. This not only has implications for understanding adults and kids but also in the context of mental and physical health. We can, for instance, ask which kinds of environments cause or exacerbate certain mental illnesses, or what types of architectural constraints lead to maladaptive behaviors. In doing so, we might be able to better understand these issues and, in turn, potentially develop targeted aids for them.

R5. Points of contention

Although the framework proposed in our target article was received well overall, there were a few points of contention raised by some of the commentators. In this section, we address and clarify these issues.

The first of them is raised by **Ong et al.** and by **Székely & Orbán**. They both argue that having to specify an inference problem is a virtue of the Bayesian approach, not a limitation. From their perspective, the process of defining the inference problem can in itself shed light on the system whose cognitive processes are being modeled. We agree that this is a valid – and often very useful – strategy. However, both of the commentaries come with the implicit assumption that this is not possible in the meta-learning framework. We think this is a wrong dichotomy. The exact same research strategy can be applied in the meta-learning framework: (1) Define a data-generating distribution, (2) draw samples from it, (3) use these to construct a meta-learned model, and (4) compare models with different assumptions against each other. To illustrate this using the probabilistic programming example of Ong et al., one could, for example, define a distribution over probabilistic programs and use meta-learning to construct a neural network that can perform approximate inference over probabilistic programs. Although we generally agree that this is a useful research strategy, it is important to mention that there are settings in which it is just not applicable, as outlined in our target article. We think this is where the strength of meta-learning lies. It allows us to do all of the things we can do in the traditional Bayesian framework – including probabilistic programs – and more.

From a conceptual perspective, we also have to weigh in on the commentaries of **Vriens, Horan, Gottlieb, & Silvetti** who state that "the framework generates models that are not interpretable in cognitive terms and, and crucially, are governed by an immense numbers of free parameters […]." That is not an accurate depiction of the meta-learning framework from our perspective. Although these models have potentially a lot of parameters, they are not free parameters that are fitted to human data. Instead, they are only optimized to maximize performance on a given task. Vriens et al. further claim that the framework generates models with low falsifiability. This is also far from the truth. Every meta-learned model can be compared against alternative models, and hence potentially falsified, as we and others have shown in many previous studies. Indeed, this is true not only on the behavioral level but also on the neural level (i.e., when there are inconsistencies with neural recordings). In this sense, meta-learned models provide even stronger grounds on which they can be refuted as pointed out by **Grant**. The meta-learning framework as a whole is of course harder to falsify. However, we believe that it should be the role of a framework to generate useful theories, not to be falsifiable.

We also noticed that there were a few misunderstandings concerning the distinction between tool and theory highlighted in our

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

target article. In particular, we put forward the notion of using meta-learning as a tool for building models of human learning. We did not say much or make any claims about how people actually acquire their learning algorithms, i.e., the process of meta-learning itself (see sect. 1.4 of the target article). Although this is an important problem, it is outside the scope of our article. **Llewellyn** states that our first question is to understand how people improve their learning abilities over time (and subsequently that we fail to address this question satisfyingly). However, we explicitly want to highlight again that we did not strive to address this question in our target article. Likewise, **Calderan & Visalli** mention that we should position ourselves against hierarchical Bayesian models, which can be used to model learning-to-learn. We agree that this would be needed if we were targeting to find out how people learn-to-learn, which we are not. Even though the study of learning-to-learn is outside the scope of our target article, we still believe that the meta-learning framework could provide an interesting perspective for studying these processes. This was, for example, noted by **Nussenbaum & Hartley** in the context of developmental psychology, or by **Yin, Xiao, Wu, & Lian** who make the connection to integrative learning.

Finally, **Calderan & Visalli** question the utility of meta-learning to build rational models in large worlds. In particular, they ask: "What justifications exist for the selection of training data?" They rightfully claim that meta-learned models have priors too and that they therefore offer no important advantages over Bayesian models. However, in contrast to Bayesian models, meta-learned models do not require an explicit expression for these priors – they only need samples from them, which is a much weaker requirement. That means that we can go out and measure them by collecting samples. In turn, this gives us many opportunities. We can, for example, ask people to generate samples from their priors (as done in the work of Kumar et al. [2022] mentioned earlier), or we can go out and determine priors that match real-world statistics (as done in the work of Jagadish et al. [2024] mentioned earlier). Meta-learning then allows us to compile these priors into a computational model. Although hierarchical Bayesian models may also be able to construct their priors, as mentioned by Calderan & Visalli, they can only do so in a predetermined class of functions, preventing an effective application to large world problems.

R6. Links to foundation models

To our surprise, none of the commentaries touched upon the similarities between meta-learning and training LLMs. We therefore wanted to use this opportunity to raise a few points on this topic ourselves. Essentially, LLMs are trained using the same objective we have discussed in our target article (equation 7). The only thing that is special is that the data distribution amounts to the whole internet. In this sense, LLMs can be viewed as a special case of meta-learned models – all the same principles apply. Thus, one way to view LLMs is that they approximate Bayesian inference to predict the next tokens in human language. Like the meta-learned models we have discussed in our target article, LLMs learn from their context (i.e., a history of previous observations) to make better predictions with more examples by updating only their internal activations. There are exciting research questions only waiting to be answered in the space between human cognition and LLMs, and we believe that the meta-learning perspective could help us in this endeavor (Binz & Schulz, 2023; Hussain, Binz, Mata, & Wulff, 2023; Yax, Anlló, & Palminteri, 2023).

It might also be interesting to think about meta-learned models that are not based on the objectives outlined in our target article. For example, we may ask how meta-learned models relate to the concepts of free energy minimization and active inference (see commentary by **Penacchio & Clemente**), what objectives are needed to meta-learn quantum models (see commentaries by **Clark** and **Mastrogiorgio**), whether we can meta-learn models using contrastive losses (Tian et al., 2020), or whether it is possible to give meta-learning systems the ability to determine their own objectives (see commentary by **Moldoveanu**). Doing so might lead to models that do not approximate Bayesian inference but have other appealing properties. Nevertheless, putting theoretical properties aside, finding out which models are most useful in understanding cognition will ultimately be an empirical question, not a theoretical one.

Where will cognitive modeling be in 10 years from now? We predict that there will be major advances in two main directions: (1) Our models will become much more domain-general and (2) they will process high-dimensional, naturalistic stimuli. The meta-learning framework will help us to achieve both of these objectives. The first is already addressed by design: Meta-learning involves training on a collection of tasks – we only have to make this collection more diverse. Regarding the second, meta-learned models of cognition can be readily combined with visual neural networks, thereby giving them the ability to "see" experimental stimuli similarly to people (as pointed out by **Sanborn et al.**). We are already witnessing some of these systems that perform a wide range of tasks in complex, vision-based environments in the machine learning literature. Examples include models such as Voyager (Wang et al., 2023), Ada (Team et al., 2023), or SIMA (Raad, Ahuja, Besse, Bolt, & Young, 2024) – all of which are based (at least to some extent) on a meta-learned model. Unfortunately, these models are currently too expensive to train for most academic research labs (let alone to run ablations on them). For example, training Ada requires access to 64 TPUs for five weeks. However, compute is getting cheaper every year, and – together with technological advances – we think it is likely that a similar system could be trained on standard hardware 10 years from now. We are excited by this prospect and what it means for understanding human cognition.

References

Anderson, J. R. (2013). *The adaptive character of thought*. Psychology Press.

Binz, M., & Schulz, E. (2022). Modeling human exploration through resource-rational reinforcement learning. *Advances in Neural Information Processing Systems*, *35*, 31755–31768.

Binz, M., & Schulz, E. (2023). Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*.

Binz, M., Gershman, S. J., Schulz, E., & Endres, D. (2022). Heuristics from bounded meta-learned inference. *Psychological Review*, *129*(5), 1042.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Buschoff, L. M. S., Schulz, E., & Binz, M. (2023). The acquisition of physical knowledge in generative neural networks. In *International Conference on Machine Learning* (pp. 30321–30341). PMLR.

Chan, S., Santoro, A., Lampinen, A., Wang, J., Singh, A., Richemond, P., … Hill, F. (2022). Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, *35*, 18878–18891.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2022). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines?. *BioRxiv*, 2022-03.

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press

Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review*, *127*(3), 412.

De Raedt, L., Manhaeve, R., Dumancic, S., Demeester, T., & Kimmig, A. (2019). Neuro-symbolic= neural+ logical+ probabilistic. In *NeSy'19@ IJCAI, the 14th International Workshop on Neural-Symbolic Learning and Reasoning*.

Griffiths, T. L., Kemp, C., Tenenbaum, J. B.. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 59–100). Cambridge University Press.

Harrison, P., Marjieh, R., Adolfi, F., van Rijn, P., Anglada-Tort, M., Tchernichovski, O., … Jacoby, N. (2020). Gibbs sampling with people. *Advances in Neural Information Processing Systems*, *33*, 10659–10671.

Hussain, Z., Binz, M., Mata, R., & Wulff, D. U. (2023). A tutorial on open-source large language models for behavioral science. *PsyArXiv preprint*.

Jagadish, A. K., Binz, M., Saanum, T., Wang, J. X., & Schulz, E. (2023). Zero-shot compositional reinforcement learning in humans.

Jagadish, A. K., Coda-Forno, J., Thalmann, M., Schulz, E., & Binz, M. (2024). Ecologically rational meta-learned inference explains human category learning. *arXiv preprint arXiv*:2402.01821.

Kumar, S., Correa, C. G., Dasgupta, I., Marjieh, R., Hu, M. Y., Hawkins, R., … Griffiths, T. (2022). Using natural language and program abstractions to instill human inductive biases in machines. *Advances in Neural Information Processing Systems*, *35*, 167–180.

Lake, B. M., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, *623*(7985), 115–121.

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, *37*(3), 243–282.

Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A., & Kornblith, S. (2022). Human alignment of neural network representations. *arXiv preprint arXiv:2211.01201*.

Ortega, P. A. (2020). Twitter. Retrieved from https://twitter.com/adaptiveagents/status/1334609817432940550?lang=bn

Ortega, P. A., Wang, J. X., Rowland, M., Genewein, T., Kurth-Nelson, Z., Pascanu, R., … Legg, S. (2019). Meta-learning of sequential strategies. *arXiv preprint arXiv:1905.03030*.

Palminteri, S., & Lebreton, M. (2022). The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in Cognitive Sciences*, *26*(7), 607–621.

Saanum, T., Éltető, N., Dayan, P., Binz, M., & Schulz, E. (2023). Reinforcement learning with simple sequence priors. *Advances in Neural Information Processing Systems*, *36*, 61985–62005.

Schubert, J. A., Jagadish, A. K., Binz, M., & Schulz, E. (2024). In-context learning agents are asymmetric belief updaters. *arXiv preprint arXiv:2402.03969*.

SIMA Team, Raad, M. A., Ahuja, A., Barros, C., Besse, F., Bolt, A., … Young, N. (2024). Scaling instructable agents across many simulated worlds. Technical Report.

Sutton, R. (2019). The bitter lesson. *Incomplete Ideas (blog)*, *13*(1), 38.

Team, A. A., Bauer, J., Baumli, K., Baveja, S., Behbahani, F., Bhoopchand, A., … Zhang, L. (2023). Human-timescale adaptation in an open-ended task space. *arXiv preprint arXiv:2301.07608*.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., & Isola, P. (2020). What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, *33*, 6827–6839.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., … Anandkumar, A. (2023). Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., … Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Yax, N., Anlló, H., & Palminteri, S. (2023). Studying and improving reasoning in humans and machines. *arXiv preprint arXiv:2309.12485*.

https://doi.org/10.1017/S0140525X23003266 Published online by Cambridge University Press